SISTEMA STATISTICO NAZIONALE
ISTITUTO NAZIONALE DI STATISTICA

Atti della Quinta Conferenza Nazionale di Statistica

Innovazione tecnologica e informazione statistica

Roma 15, 16, 17 novembre 2000

Programma

15

novembre 2000

09,00	Arrivo ed iscrizione dei partecipanti
	SESSIONE PLENARIA
10,00	Saluto delle autorità
10,30	Innovazione tecnologica e informazione statistica Alberto Zuliani
11,00	*Intervallo*
11,30	Inaugurazione del salone dell'informazione statistica
13,00	Ricevimento di benvenuto
	SESSIONE PLENARIA
14,30	Seizing opportunities to inform the information society: creating value from statistics through exploiting new technologies Len Cook
15,00	Data mining and official statistics Gilbert Saporta
15,30	Dibattito
16,30	*Intervallo*
16,50	Presentazione di prodotti e realizzazioni
19,00	Chiusura dei lavori della prima giornata

Durante i lavori della prima giornata sarà assicurata la traduzione simultanea italiano - inglese

16

novembre 2000

09,00 SESSIONI IN PARALLELO
Integrazione tra registri e rilevazioni statistiche
Gianni Billia
Basi informative integrate per l'analisi dei fenomeni sociali
Lorenzo Bernardi
Telerilevamento e informazione statistica
Giovanmaria Lechi
Esperienze di riorganizzazione di processi statistici nel Sistema statistico nazionale
Luigi Marini

11,00 *Intervallo*

11,20 TAVOLA ROTONDA
Tecnologie, conoscenze, competenze
Federico Butera, Sebastiano Bagnara, Giampio Bracchi, Alessandro Osnaghi, Linda Laura Sabbadini

Presentazione di prodotti e realizzazioni

13,30 *Intervallo*

Programma

16

novembre 2000

14,30 SESSIONI IN PARALLELO
Il progetto SAIA, un'opportunità e una sfida per il Sistema statistico nazionale
Roberto Benzi
Sistemi distribuiti e architettura dei dati
Giorgio Ausiello
Data warehousing in ambiente statistico
Federico Rajola
Sistemi informativi territoriali
Cristoforo S. Bertuglia

16,30 *Intervallo*

16,50 TAVOLA ROTONDA
Statistica e società dell'informazione
Alessandro Penati, Giovanni Barbieri, Elio Catania, Roberto Galimberti, Raffaele Simone

Presentazione di prodotti e realizzazioni

17

novembre 2000

09,00 SESSIONI IN PARALLELO

Connessione in rete, sicurezza informatica e riservatezza
Antonio Lioy

Tecnologia web per la produzione statistica
Salvatore Tucci

Editoria elettronica per la diffusione dell'informazione statistica
Donato Speroni

Problemi di misurabilità della società tecnologica
Franco Malerba

11,00 *Intervallo*

11,20 SESSIONE PLENARIA

Relazione del direttore del Dipartimento della segreteria centrale del Sistema statistico nazionale
Vincenzo Lo Moro

Relazione del presidente della Commissione per la garanzia dell'informazione statistica
Ugo Trivellato

Relazione del direttore generale dell'Istituto nazionale di statistica
Giuseppe Perrone

13,00 Conclusioni

Innovazione tecnologica e informazione statistica

Roma 15, 16, 17 novembre 2000

INNOVAZIONE TECNOLOGICA E INFORMAZIONE STATISTICA

Alberto Zuliani

presidente dell'Istituto nazionale di statistica

If nature has made any one thing less susceptible than all others of exclusive property, it is the action of the thinking power called an idea, which an individual may exclusively pos - sess as long as he keeps it to himself; but the moment it is divulged, it forces itself into the possession of everyone, and the receiver cannot dispossess himself of it: its peculiar character, too, is that no one possesses the less, because every other possesses the whole of it
(Thomas Jefferson).

Rapidità del cambiamento

La tecnologia sta cambiando i modi di produrre beni e servizi, le abitudini, i consumi, l'organizzazione del tempo, l'intero tessuto della società. La velocità con la quale il cambiamento si produce ci ha colto, in larga parte, di sorpresa. Internet è nata come rete di comunicazione a fini militari, ridondante e quindi meno vulnerabile; per lungo tempo, ha costituito un canale tra università e centri di ricerca. Il concetto di world wide web e i programmi che lo hanno reso praticabile sono nati non presso i grandi produttori di hardware o software ma, quasi per caso, a opera di singoli "visionari" e in luoghi in cui non costituivano l'obiettivo primario. Lo standard Html è stato elaborato nel 1989

da Tim Berners-Lee al Cern di Ginevra e Marc Andreessen ha lanciato Mosaic, il primo browser, all'inizio del 1993 quando era studente all'Università dell'Illinois.

Tuttavia, la creatività dei singoli ha avuto l'opportunità di immergersi in un contesto di rete, anche informale, un tessuto condiviso che ha consentito di contrastare la dispersione delle idee e conseguire risultati di valore e finalità diverse rispetto agli obiettivi iniziali.

Tratteggiare uno scenario anche prossimo è quindi un esercizio tutt'altro che banale. L'innovazione procede a volte per discontinuità, anche se l'affermazione che la natura e la cultura non fanno salti rimane in genere valida. Accade così che giunga inaspettata e ci spiazzi, soltanto perché non stavamo guardando nella direzione giusta, non guardavamo le cose giuste nonostante talvolta fossero accanto a noi.

Molte altre tecnologie sono potenzialmente pervasive: le biotecnologie, l'ingegneria genetica, la nanotecnologia. In ciascuna di esse è presente, in misura più o meno implicita, una forte componente di tecnologia dell'informazione. Ciò aumenta la complessità e rende ancora più arduo prevedere le conseguenze.

Società dell'informazione

Queste difficoltà vanno tenute presenti se si vuole riflettere sulla relazione tra innovazione tecnologica, soprattutto tecnologie dell'informazione, e informazione statistica.

La componente informatica è ovviamente importante, poiché consente di trattare grandi volumi di dati e di trasmetterli in modo rapido e sicuro. Ma soprattutto è rilevante il modo in cui l'informazione deve essere organizzata per consentirne l'integrazione, un valore emergente anche in campo statistico.

L'informazione è la cosa più simile al pensiero che sia mai esistita: la sua espressione si avvia a diventare indipendente dal supporto fisico; la distinzione tra originale e copia cessa di avere significato; gli oggetti fisici su cui risiede in forma digitale non hanno rilevanza, né valore, in quanto tali; chi la utilizza non ne entra in possesso nel senso tradizionale del termine. Forse è bene che sia così, perché se l'informazione è simile al pensiero e rappresenta in misura crescente una fonte di reddito e ricchezza, ostacolarne la libera circolazione rischierebbe di frenare lo sviluppo.

Altre caratteristiche sfuggono alle categorie che si applicano alle merci e sulle quali si esercitano le leggi dell'economia e le regole che ne governano l'ordinato funzionamento.

In primo luogo, l'informazione si comporta come una forma di vita. Il biologo inglese Richard Dawkins ha proposto quasi venticinque anni fa l'idea

di “meme”, l’unità minima di informazione capace di duplicarsi e di propagarsi negli eco-sistemi della memoria. I memi, come i geni, si riproducono, interagiscono con l’ambiente, si adattano, occupano nicchie, persistono. Analogamente, la trasmissione digitale dell’informazione, come quella orale, favorisce la sua continua evoluzione, cosicché non esiste mai un’edizione definitiva. D’altra parte, proprio per questo, l’informazione tende a essere altamente deperibile, anche se la durata della sua vita utile dipende in larga misura dalla sua natura e dagli scopi dei fruitori. Quella statistica non fa eccezione.

In secondo luogo, l’informazione è definita meglio come attività che come oggetto, si manifesta come un’interazione tra menti umane. Essa può essere bensì immagazzinata in un supporto fisico, ma esiste soltanto in una transazione; per esprimersi deve essere “decompressa” e sperimentata da chi ne è destinatario. Partecipa anche largamente della natura dei beni pubblici.

In terzo luogo, l’informazione si manifesta all’interno di un rapporto sociale. Il suo significato si realizza nel luogo e nel momento in cui la trasmissione diviene ricezione e dipende essenzialmente dalla disponibilità di recettori adeguati: attenzione, interesse, linguaggio, terminologia, paradigmi condivisi.

La consueta relazione inversa tra scarsità e valore, così importante per i beni materiali, in genere non sussiste; al contrario, in molti casi, sono la familiarità e la reputazione a dare più valore a un’informazione. Operano in questa direzione due meccanismi. Il primo ha a che fare con la creazione di standard di fatto, allorché più persone comunicano utilizzando il medesimo sistema di convenzioni e di informazioni (un linguaggio, un paradigma scientifico, una base di dati). L’importanza di una rete cresce dunque in ragione del numero di interconnessioni possibili. Il secondo ha a che fare con l’autorevolezza della fonte: in una situazione in cui l’informazione è sovrabbondante, viene attribuito valore maggiore a quella, pertinente e di buona qualità, proveniente da un soggetto che in passato abbia dimostrato di essere attendibile. Questo determina la sua ulteriore diffusione.

In definitiva, l’informazione va intesa come un servizio fondato sulla possibilità di un accesso interattivo del fruitore alla competenza del produttore.

Dal punto di vista della statistica ufficiale, ciò comporta ulteriori investimenti in tre direzioni: la qualità, anche attraverso le possibilità di integrazione, e la trasparenza dei processi produttivi; la pertinenza e l’interpretabilità dei contenuti dal punto di vista degli utilizzatori; la certezza dei tempi di rilascio.

Le tecnologie dell’informatica e delle telecomunicazioni – e segnatamente Internet – pongono, inoltre, nuovi problemi con riferimento ad almeno due aree di grande rilevanza:

- la privacy, anzitutto: ai rischi che la presenza di vaste raccolte di dati personali pone in generale – soprattutto con riferimento alla possibilità di loro uso incontrollato, alle potenzialità di monitoraggio e di controllo dei comporta-

menti dei singoli, alla sicurezza e all'autenticità delle transazioni – si aggiungono responsabilità specifiche per gli operatori della statistica ufficiale, per le potenzialità informative di patrimoni tanto estesi, le quali possono determinare nuove asimmetrie;

- la selezione dell'informazione e la sua censura che ne è l'aspetto estremo: di fronte a giacimenti informativi vasti, la riduzione della complessità e la ricerca di pattern dotati di significato sono un compito specifico del lavoro statistico, che si è evoluto dalle tecniche più tradizionali a quelle di data mining; c'è il rischio, tuttavia, che questi processi possano condurre a una selezione capace di distorcere l'informazione ed esso trascende il terreno della correttezza metodologica per investire quello della deontologia. Le semplificazioni operate dalla statistica ufficiale debbono sempre rappresentare un servizio, inteso alla documentazione e all'analisi dei fenomeni. Se fossero percepite come una manipolazione della realtà, guidata da valutazioni soggettive non giustificate o se il processo non fosse trasparente, verrebbe meno la fiducia che è alla base del patto sociale su cui si fonda la statistica ufficiale. La configurazione del sistema italiano garantisce, tuttavia, da pericoli di questo tipo.

Questi problemi non ammettono risposte facili o sottovalutazioni. Sono certamente necessarie garanzie, ad esempio, sul rispetto della privacy, sulla tutela del segreto statistico, sugli standard applicati alla elaborazione dei dati. A queste regole deve aderire il sistema statistico nel suo complesso, nelle articolazioni del Sistema statistico nazionale e nell'ambito della cooperazione internazionale. Soltanto la condivisione di intenti, valori e criteri deontologici può garantire autorevolezza alla statistica ufficiale: in un mondo dell'informazione "globalizzato", nessun attore e nessuno Stato possono pensare di poter agire isolatamente.

Altrettanto necessario è il miglioramento della capacità di valutazione da parte degli utilizzatori; su di essi graverà inevitabilmente una parte crescente degli oneri di vigilanza e controllo sull'informazione disponibile in rete. Dovranno essere in grado di discernere l'autorevolezza delle fonti, anche diverse da quelle certificate, così che possano essere scoperti e valorizzati nuovi giacimenti informativi; dovranno saper valutare i limiti di validità delle informazioni; dovranno saper orientarsi e discriminare tra informazioni in conflitto, magari soltanto apparente; dovranno maturare un'autonoma capacità interpretativa. In questo processo, non potranno essere lasciati soli: la società dell'informazione richiede un forte sviluppo della cultura statistica, e in generale della numeracy; dunque, un investimento formativo per il quale la scuola e la statistica ufficiale dovranno cooperare strettamente.

Verso una società della conoscenza

Il riferimento all'informazione come servizio contribuisce a chiarire le condizioni del passaggio dalle tecnologie dell'informazione alla società della conoscenza. L'utilità percepita dell'informazione, la pertinenza commisurata ai bisogni conoscitivi, la possibilità di impiegarla per assumere decisioni, riducendo l'incertezza e la complessità, rappresentano requisiti preliminari per la sua trasformazione in conoscenza. Schematicamente, la transizione è dal dato (rappresentazione di un fenomeno), all'informazione (comprensione del dato attraverso l'interazione tra soggetti sociali e l'attribuzione di significato da parte del destinatario), alla conoscenza (riconoscimento di valore all'informazione, in funzione del soddisfacimento di un bisogno ed eventualmente dell'assunzione di una decisione).

In realtà, il processo è più articolato. L'informazione, per trasformarsi in conoscenza, deve essere interiorizzata, tanto dagli individui quanto dai soggetti sociali. Deve potere interagire con altra informazione. Deve essere utilizzata all'interno di processi che creano "metafore", per applicarla a problemi e ambiti diversi rispetto a quelli per i quali era stata originariamente prodotta.

Il carattere sociale e sistemico dei processi conoscitivi pone molti problemi a chi ha la responsabilità, in quanto operatore pubblico, di alimentare la conoscenza e di fare in modo che tutti abbiano un accesso adeguato all'informazione.

In primo luogo, c'è da chiedersi se le opportunità offerte dalle tecnologie dell'informazione siano equamente distribuite. I dati disponibili, riferiti al 1999, collocano al 29 per cento la percentuale di famiglie in possesso di uno strumento informatico. Soltanto il 20,9 per cento risulta in possesso di un personal computer. La quota scende al 10,5 per cento fra quelle in cui la persona di riferimento non è in condizione professionale, sale al 31,4 per cento fra le altre, raggiungendo punte del 52,3 per cento per dirigenti, imprenditori e liberi professionisti.

Sono ancora forti gli ostacoli di carattere culturale: le percentuali più alte di persone che utilizzano il personal computer o accedono a Internet si registrano nelle classi di età comprese tra 14 e 44 anni; per quelle successive si abbassano bruscamente.

D'altronde, fattori di discriminazione simili operano anche per l'accesso tradizionale all'informazione, ad esempio attraverso i quotidiani.

Fenomeni della stessa natura interessano le imprese: sempre nel 1999, soltanto il 22,2 per cento si avvaleva di tecnologie informatiche. Sopra i venti addetti, tuttavia, le percentuali salgono repentinamente, al 75 per cento per quelle con 20-49 addetti, al 90 per cento per quelle medie (50-249 addetti), al 96 per cento per le grandi (250 addetti o più). Anche per le imprese sembrano operare fatto-

ri culturali in senso lato: quelle di minori dimensioni mostrano un'attitudine maggiore ad avvalersi delle tecnologie informatiche per la commercializzazione e la progettazione; all'aumentare della dimensione aziendale assume importanza crescente l'utilizzazione per la produzione e la movimentazione. Ciò è coerente con le strategie differenziate che qualificano rispettivamente i settori industriali caratterizzati dalla grande e grandissima dimensione, dove hanno preminenza gli aspetti tecnologici legati alla ristrutturazione delle produzioni su grande scala, e quelli dell'industria leggera, punti di forza della specializzazione produttiva nazionale, caratterizzati dalla media e piccola dimensione, dove assume rilievo l'attenzione al mercato e alle relazioni con i clienti.

Un altro tema stimolante è quello della misurazione di un'economia e una società della conoscenza. Vi è, anzitutto, il problema della quantificazione dell'informazione che si produce sulla rete: ad esempio, per le transazioni dell'e-commerce e del trading on-line, molti concetti e convenzioni della statistica ufficiale non valgono più (a quale spazio e a quale tempo ha senso riferire la transazione?); in generale, i processi della new economy sono immateriali e di difficile documentazione statistica.

Se si considera l'aspetto della conoscenza come informazione utile, occorrerebbe misurarla in termini di decisioni assunte sulla base dell'informazione sottostante; una linea, però, difficilmente percorribile nel concreto. Una prospettiva interessante è aperta dalla considerazione che il passaggio dall'informazione alla conoscenza risiede nella possibilità di riutilizzarla in contesti decisionali e operativi diversi: ciò è alla base delle statistiche sugli accessi a una banca dati o un sito Internet.

Sistemi informativi

Queste considerazioni conducono a un ulteriore tema, di particolare importanza per chi opera nell'ambito della statistica ufficiale: se l'informazione si traduce in conoscenza, in virtù della possibilità di sua utilizzazione integrata in contesti decisionali diversi da quello che originariamente ne aveva stimolato la produzione, qualunque strumento operativo atto a favorire processi di questo tipo assume grande importanza.

Si impone così il paradigma di sistema informativo. Da una parte, le potenzialità rese disponibili dai progressi in tema di database relazionali e in generale i risultati dell'attività di ricerca in campo informatico hanno contribuito alla riorganizzazione del ciclo produttivo delle indagini; dall'altra, gli sviluppi di metodologie e tecniche di integrazione delle fonti consentono di avvalersi, sempre più frequentemente e utilmente, di dati amministrativi. In un modello integrato, una pluralità di fonti e una pluralità di prodotti informativi

sono mediati (concettualmente e operativamente, quindi sotto il profilo statistico e informatico) dal sistema informativo.

A questo approccio, diversi apparati nazionali di statistica stanno lavorando da tempo, consapevoli, d'altronde, che gli utilizzatori si pongono sempre più spesso in un'ottica di ricerca dell'informazione d'interesse sfruttando una pluralità di giacimenti informativi, con contenuto e caratteristiche differenti. Ridefinire il prodotto dell'attività statistica come sistema informativo richiede che, per ciascuna delle fonti, siano documentate le caratteristiche rilevanti; in altri termini, che sia resa disponibile la "meta-informazione".

Dal punto di vista organizzativo, assumono grande importanza le procedure di accesso all'informazione, concepita come un servizio provveduto con continuità a utenze anche non bene specificate.

Modi nuovi di produrre informazione

Nei settori di punta dell'economia, in particolare nell'industria del software, l'innovazione tecnologica e organizzativa lascia intravedere un modo di produrre che mette in secondo piano l'approccio basato sulla progettazione puntuale e gerarchica dei passi del processo, a favore di uno in cui singoli oggetti e componenti vengono assemblati per tentativi ed errori, con larga autonomia e spazio alla creatività. Sono definiti soltanto l'architettura generale e le funzioni che il prodotto dovrà svolgere, selezionate sulla base delle esigenze e dei bisogni degli utenti finali. Gli standard organizzativi e tecnologici non sono progettati ex ante, ma emergono in funzione della loro performance in situazioni operative differenti e della reputazione guadagnata sul campo. È un modo di produrre che tende ad esaltare la fertilizzazione tra competenze relative a campi diversi e a favorire l'emergere e l'affermarsi dell'innovazione.

Nel nostro settore, l'attività di definizione dei contenuti delle indagini e dei sistemi informativi, che precede e orienta la realizzazione dei processi operativi, è non pienamente standardizzabile e largamente basata sul contributo dei ricercatori. Viceversa, è tradizione perseguire un elevato grado di standardizzazione riguardo al trattamento dei dati, condizione per la ripetibilità e documentabilità. Un ambito nel quale l'approccio "destrutturato" appare coerente con la dinamicità del cambiamento è quello della distribuzione e dell'uso dell'informazione prodotta. Tradizionalmente, le organizzazioni della statistica ufficiale hanno mediato tra una pluralità di fonti e di processi produttivi e le esigenze dell'utenza attraverso un'interfaccia semantica tra il "giacimento di informazione", rappresentato dal medium fisico ove risiedono e sono organizzati i dati provenienti dalle diverse fonti, e i bisogni informativi. La progressiva articolazione di questi, il moltiplicarsi delle fonti d'informazione, la crescente com-

plessità dei processi produttivi comportano requisiti nuovi per l'interfaccia, che deve consentire all'utente la definizione della propria "vista" sull'informazione. In questa direzione si muovono le prospettive del data mining, cui è stato recentemente dedicato un interessante convegno qui a Roma e sulle quali si soffermerà nel pomeriggio il contributo di Gilbert Saporta.

Sotto un profilo diverso, toccato dalla relazione di Len Cook, la tecnologia web può contribuire ad arricchire ulteriormente l'interfaccia semantica, attraverso la possibilità di far transitare sulla rete (nelle sue componenti intranet ed internet) la gestione e la documentazione dei processi e di rendervi disponibili strumenti per produrre dinamicamente contenuti informativi specifici.

Opportunità di crescita per il Sistema statistico nazionale

Le tecnologie dell'informazione consentono maggiore libertà organizzativa. Si modificano perciò i termini tradizionali della tensione tra centralizzazione e decentramento in campo statistico: in passato, il luogo in cui i dati risiedevano – ed era economica ed efficiente la loro centralizzazione – condizionava fortemente il sistema, limitando l'autonomia per la sua progettazione e gestione e le stesse opportunità di accesso da parte degli utilizzatori; oggi e ancora più nella prospettiva del data mining e dell'impiego integrato delle tecnologie web, il luogo fisico in cui i dati sono prodotti o risiedono perde di rilevanza, a patto che le procedure di accesso siano garantite e facilitate a tutti i livelli e l'interfaccia semantica sia messa a disposizione di una pluralità di utenti e di gestori.

La riforma delle autonomie e i processi di valorizzazione in atto sono guidati dall'intento di avvicinare le decisioni sui servizi pubblici quanto più possibile ai cittadini che li utilizzano, salvaguardando l'efficienza economica. Gli obiettivi sono quello di rendere più lineare il percorso che va dall'individuazione dei bisogni all'erogazione dei servizi e quello di responsabilizzare il decisore pubblico nei confronti dei cittadini. Per ambedue, è necessario disporre di informazione statistica. Il luogo di produzione di quest'ultima perde di rilevanza, salvo che venga generata contemporaneamente all'erogazione del servizio. Rimane essenziale che risponda a standard omogenei, per garantirne la comparabilità, in modo da favorire l'individuazione e la misurazione dei bisogni e il "governo" delle politiche e da consentire la valutazione dei risultati da parte dei cittadini e di terzi.

La possibilità di rinvenire presso altri soggetti contenuti informativi e moduli dei processi di produzione, interrogazione e diffusione, da importare nel proprio ambiente e utilizzare nella propria situazione, può essere di grande stimolo alla crescita del Sistema statistico nazionale.

Nuove funzioni statistiche

L'interfaccia semantica non può ridursi a un insieme di procedure informatiche, ma richiede l'intervento di intermediari, capaci di interpretare le esigenze di utenti non sempre attrezzati a esplicitarle, di fornire strumenti concettuali idonei a individuare le informazioni specifiche che possono sostenere i processi decisionali, a chiarirne i limiti e gli ambiti di applicazione.
Gli statistici devono essere protagonisti dell'information brokering. Un primo salto di qualità è stato fatto passando dalla cultura del giacimento di dati alla cultura dell'informazione orientata all'utenza. Ci si deve spingere più avanti, divenire artefici del suo utilizzo.

Gli statistici possono diventare attori anche del knowledge brokering, cioè di un approccio alla conoscenza tradotto in chiave organizzativa. La condivisione, l'integrazione, la trasmissione e la fertilizzazione delle idee in prospettiva sociale non possono avanzare soltanto per effetto di spinte tecnologiche. Gli statistici ufficiali devono svolgere un ruolo importante, agevolando la condivisione dell'informazione e la sua applicazione a contesti diversi; d'altra parte, la stessa organizzazione del Sistema statistico nazionale può avvalersi della modularità offerta dal knowledge brokering, ad esempio, in tema di trasferimento di tecnologie e strumenti e di condivisione di basi di dati e di sistemi di meta-informazione.

Conclusioni

Le tecnologie dell'informazione hanno alterato, e ancora più lo faranno, assetti e dinamiche organizzative consolidate. In un sistema nel quale si moltiplicano le fonti, la risposta della statistica ufficiale non può essere quella di sanzionare giuridicamente la propria produzione ma di guadagnare autorevolezza attraverso la qualità riconoscibile e, in primo luogo, interpretando correttamente i bisogni e fornendo elementi di orientamento e valutazione riguardo all'informazione fornita. Poiché aumenterà il carico di responsabilizzazione dell'utenza, essa dovrà anche essere attrezzata per discernere.

Le tecnologie offrono interessanti opportunità per l'ulteriore edificazione del Sistema statistico nazionale. Le più interessanti sono legate alla condivisione dell'informazione e alle possibilità di accesso. Il decentramento delle funzioni decisionali e amministrative resta un tema importante sotto il profilo istituzionale, lo è meno sotto quello informativo; i valori da perseguire sono l'adozione di standard comuni, l'integrazione delle fonti e dei sistemi informativi, l'accesso ai dati che sarà agevolato dall'introduzione del codice deontologico.

In presenza di processi di cambiamento tecnologico intensi e veloci

come quelli in atto, c'è il rischio che le differenze si accentuino e possano ostacolare la costruzione del Sistema. Occorre quindi mantenere un'attenzione continua alla propagazione delle soluzioni metodologiche e tecniche e soprattutto condividere un'identità di scopi che poggi su alcuni motivi profondi: la consapevolezza di poter svolgere un ruolo importante nell'evoluzione verso la società della conoscenza; conseguentemente, una finalizzazione dell'informazione agli utilizzatori, predisponendo e rendendo disponibili strumenti per ricomporre complessità e ricchezza di opportunità, con la legittima aspettativa di un sempre più ampio riconoscimento della funzione statistica nel paese.

Riferimenti bibliografici:

Barlow, John Perry (1994), "The Economy of Ideas", Wired 2.03.
Barlow, John Perry (2000), "The New Economy of Ideas", Wired 8.10.
Blackmore, Susan (1999), The Meme Machine, Oxford University Press, Oxford.
Dawkins, Richard (1976), The Selfish Gene, Oxford University Press, Oxford.
Dyson, Esther (1994), "Intellectual Property on the Net", Release 1.0.
Dyson, Esther (1997), Release 2.0: A Design for Living in the Digital Age, Broadway Books, New York.
Zuliani, Alberto (1996), Informatica e ricerca nella statistica ufficiale, Scuola superiore Sant'Anna, Pisa, 8-9 novembre 1996 (mimeo).

SEIZING OPPORTUNITIES TO INFORM THE INFORMATION SOCIETY: CREATING VALUE FROM STATISTICS THROUGH EXPLOITING NEW TECHNOLOGIES

Len Cook

Office for National Statistics, United Kingdom

Introduction

Statistical data is very important to public policy, and to public understanding of the context of public policy. Its accessibility contributes to the effectiveness and trust in government, and the efficiency and stability of financial, commodity, labour and other markets. The effective delivery of official statistics on the internet is not only driven by the potential efficiencies for continuing to deliver existing services to existing users, but the growing expectation of access at a low cost by citizens to government services, higher levels of numeracy and a sea change in the ways in which statistics are able to be made interesting and relevant. Our future service model may well come from the media, compared the present more academic-like approach we take.

The internet is a fast growing, highly integrated global facility, and the competitive position of the statistical office internet service, its comparative qualities, and the nature of relevant innovations, are highly visible through the being provided alongside other public services, and alongside those of all other information providers, in each and any of the situations users are in when they draw on our services.

The special value of the internet to official statistics

The internet is an architecturally independent pathway that enables myriads of linkages through public and private communications networks between a large minority of private households, many businesses, most public bodies, many community organisations and nearly all of the media, research and international organisations. The internet enables a gigantic leap in the numbers of people who can access official statistics at a very low cost, regardless of region, physical mobility or language.

The internet has particular value to official statistics, because of the capacity to present statistical information dynamically, in visual form through graphs and video, through real-time linkages to databases, and linked to contextual information provides for a major step change from the static nature of books, whether in printed form or acrobat file. The ability to make statistics understandable, interesting and relevant can increase greatly, at little cost.

Through the power of contemporary web based tools in providing content sourced from databases, the internet enables powerful searching of information with or without structure, making accessible huge repositories of information perhaps originally organised in ways that were only relevant to the uses that the information was initially designed to be used for. Users can easily gain familiarity with what is available, and readily retrieve it, and learn of whether similar material exists, with little time or skill. The users' desktops can be customised to fit their interests, in the mapping, searching and presenting of information.

The internet can enable information to be efficiently provided in the great variety of forms relevant to official statistics, particularly text, statistical tables, graphs, maps. Such releases can be made available in complete studies, as tightly focused pictures of the state, organised to whatever extent is necessary to make a reasonably complete parcel of information.

The internet can integrate data, voice, visual and other means of interaction.

The internet enables each and every communication to be concluded in a satisfying way, across the huge variety of circumstances through which users of statistics engage with official statistics, regardless of how it was initiated. It provides for the resulting service delivered to users to be able to be customised to their needs, and ability to absorb statistical results.

Monitoring of activities of the statistical office web-site, including degree and form of access, enables patterns of behaviour, levels of use, and reactions to change to be cheaply and quickly assessed, at the level of the system, the specific service, and the individual user. Changes in demand, and the value placed on new services can be quickly and accurately measured, in aggregate, and for individual or groups of users.

The internet enables a high level of integration with related services, including billing and promotion of new material. It also facilitates the efficient, downstream production of internet based services by added value information providers

The nature of the statistical office web-site

The statistical office web-site has the potential, among other things, to become the prime means of delivering official statistics, managing statistical enquiries and meeting obligations to provide access to information about statistical practice. All statistical office services could be designed to service web based access, and other forms of access will be efficient by-products of the web service and associated facilities, with few exceptions.

The statistical office web-site could be a wholly integrated part of internal statistical office statistical and business systems. Information will be organised so that the power of the web tools and browsers for unstructured searches can compliment any structured access paths. The web focus of all internal processes will not usually need any intervention in business processes to provide for the results to be delivered on the web, and any will be, of necessity, quite limited.

Statistical material and other information and web-site management will be through presentations that are designed to appear firstly as quality presentations on an internet screen. Where printed, or where delivered in machine readable form, such as excel spreadsheets, that will determine the design of the information.

A typical web-site might consist of five quite distinct elements, which will be linked where relevant.

1. Pictures of the nation: These will be a series of graphical and text based presentations with dynamic elements, which present self contained and usually short stories about particular topics. These will be dynamically updated, where they are focused on the latest available measures. They may contain videos, audios, graphics and other tools embedded in them, that provide a dynamic and interactive element. There will be a high editorial contribution from the statistical office, including design and selection not only of measures but how they are presented. Particular elements will be

- The analysis and main results associated with each first release of statistics
- The nation's yearbook in a new form
- Topic reports that reflect the evolution of existing services such as eco-

nomic trends, social trends, labour trends
- Special analytical reports
- High profile statistics
- Major new statistics (i.e. Census 2001)

2. Information gathering: This will provide structured and unstructured access to:
- the formal publications and presentations of the statistical office, as well as available but unpublished tabulations and series.
- the structured statistical data-banks necessary for making available at low cost repeated access to indexed, machine readable data, particularly time series, regional statistical aggregates and accompanying aerial descriptions, and multi-way cross tabulations, most of which are based on a few critical cross cutting variables such as family, industry, sector, commodity, population group, age or sex.
- Statistical summary data-files in special areas particularly foreign trade, where a well defined set of analyses of individual import or export commodities exists.

3. Secure access to confidential unit record statistical data sources: Retrieving information from survey and administrative data sources:
- Statistical unit record database extractions, where users can define tables or other aggregates to be extracted from unit record databases using flexible self specified enquiries, and receiving results that are tested against confidentiality protection rules.
- Modelling environments, which may be made up of a mix of unit records, statistical summary information or other statistics, and to which particular models can be applied, or which are linked to genera lised modelling environments such as SAS.

4. Documentation: This will provide structured and unstructured access to:
- Writing about statistical practice, statistical processes, methods and tools. It will also provide access to meta-data of all forms, including classifications, definitions, frameworks such as ESA,
- Information about each statistical survey
- Information about the statistical office
- Information about National Statistics
- Catalogue of services

- National statistics strategies and work programme
- Links to international and other relevant country web-sites

5. Special services: These will be special services that develop for particular user communities, focus on particular products, or increase visibility of something important. Examples are;

- Education services
- Communications channels (feedback, email, voice messages, trigger for personal contact)
- Media services

Distinct faces of the statistical office web-site

The statistical office web-site may need to have two quite distinct faces, each of which will present a distinct menu of the available statistical office internet services.

1. A public face, which will be focused on immediately usable material, free access, supported by a fixed amount of telephone support each year. All documentation will be available, and several information gathering services will be available even where chargeable, so that users can learn of their existence. The public face will be designed to need no training for its use. The monitoring of each user's access will enable some customising of feedback to users, and we could plan to have a variety of front ends to the web-site, if distinct groups of users had specific and particularly different needs that justified that. There will be an integration of voice with data.

2. A "subscriber managed" face, that will provide services that are either needing access to confidential records held only in statistical office custody, or which use services that are provided on a subscriber basis, or which need training and support for their effective use, such as the time series access service. The needs of each and every user of this service will be of sufficient value to the statistical office, or of sufficient importance to government, that they might justify their own access path through the web-site services, as well as their own approach to linking to parts of the web-site content either on their own internal network, through the internet or through some private network. The integration of the statistical office web-site with other services used by these users could be customised to their needs. Training will be available for servi-

ces that need it. Security will be comprehensive, and user specific, using civil service wide infrastructures where available. There will be an integration of voice, data and video. A chargeable service enabling access to statistical office experts will be available.

Having two distinct web faces will ensure that the public face is consistent with free access to statistics, has the highest level of consistency in the look and feel of its services, has minimal security overhead, and is reproducible in many ways. The subscriber managed face will firstly be designed to deliver to policy ministries a comprehensive, easy to use, always available and well supported access to statistical information to ensure that all relevant information is discovered and accessible, for whatever question is being addressed. Other specialist users outside government will also use these specialist services, which will be more limited for unit record information except where there is the same legal authority to have access in the form potentially available.

Integration with production processes

The statistical office web-site will be fully integrated with statistical office systems and processes. Document management in some sites such as Statistics New Zealand is managed through mirroring the information provided in a Lotus Notes based knowledge management system, so that the marginal costs of transferring information into the web environment includes no translation cost. The preparation of statistical pictures should involve tools that belong to the statistical office standard tool-kit, and which by design integrate at no extra cost into the web environment. This might include Microsoft tools such as excel, or necessitate a new standard graphing tool. The dynamic access to statistical measures needs contemporary web integratable summary data management tools for time series (Fame is a good example), regional data (Supermap is a good example), and multi-way tables (Supertable or IBM Pivot table are good examples). Statistical office tools in this area may well all need upgrading.

The information management capability should accept enquiries on statistical records driven by web managed tools, and should provide extractions of data in a form that readily inputs into the standard web integrated summary data management tools.

There should be no web content that requires operator intervention or special processes to translate the information onto the web-site. The consequence of this is that the statistical office web-site will be developed in parallel with Information management initiatives.

Web based improvements in the presentation of official statistics

The web creates the potential for the presentation of official statistics to radically improve in quality in several ways. Each of these will challenge the existing practices of the statistical office, but also will generate more added value from existing major work such as the studies that make up the UK Social Trends. We will see:

1. Timeliness improvements
2. Quick extraction by users of indexed, machine readable statistical series, and of graphics
3. Greater emphasis on graphics, and maps
4. New forms of presentation
5. More relevant integration of information
6. Dynamic response when changing the cross cutting variables of any analysis
7. Ease of quick linkage to related sources of information, and to relevant statistical services
8. Easy links to meta data associated with any series
9. Easy fit of results of enquiries with other electronically provided outputs

Implications of a centralised web publishing service

The statistical office information technology architecture will embrace the internet, creating a wholly integrated environment. This integration covers technology, security and information, involves tools, systems, processes, management, client, suppliers, reporting. We will create high value from communicating via databases, and have efficiencies gained by structuring information that was previously unstructured. There will be a very large web site with authorship decentralised but control centralised. There will be considerable ease of exchange of information with partners who use notes. The statistical process and data access benefits we will obtain from the fully web integrated architecture will be

1. Extensive public documentation at low cost
2. Hugely improved performance monitoring
3. Lower cost structure for all activities
4. Reduced cycle times for new activity
5. Credible capacity for more innovation

6. More rapid creation of new products
7. Able to manage more change at any time
8. Web can provide several channels for supply and delivery of services
9. Much reduced communication costs
10. Easy links to any relevant organisation
11. All statistical office statistical services are able to be digitised
12. Web compatible services penetrating down into all business processes
13. Exchangeable business systems and processes

The new statistical office web-site will be a statistical office wide facility, relevant to all statistical outputs and services, and all statistical outputs and services will use it, as the prime means of service delivery.

This will require strong corporate standards that are relevant to all the work of the statistical office, state of the art corporate systems that are used locally, everywhere, and the application of corporate practices at all times.

To facilitate this, the integration of technology environments will necessitate a high degree of standardisation of systems and equipment, integrated software and data management tools, one statistical office publishing policy, and one release process.

We need to determine our corporate tools, e.g. Lotus Notes, and apply universally without question. A strategy for settling on new publishing approaches for statistics needs to be prepared. We need to decide on what sort of books and other printed material we will continue with, to balance what the web can now best deliver.

The new statistical office web site will be tightly integrated with the new statistical office information management and knowledge management tools, and the developments will be planned and implemented in parallel.

Conclusion

For any statistical office, the web-site is a critical initiative. We need some clear direction on its focus, integration with statistical office infrastructure, and the potential changes that it will make possible.

The technology options are considerable, and when we have succeeded in them, we will find that business change may be slower than that of the technology. New technologies that exist now give statistical offices the chance to be at the forefront of public sector modernisation, provided we rethink; and redevelop the statistical services we provide.

The internet will be fully exploited by the national statistical office

when there is a civil service wide secure network to work within, all agencies have a similar level of competence in IT matters, and public access to the internet is judged equitable.

This paper is about what we can achieve with or without these broader successes, and about the contribution we can make to making these successes visible when they occur.

DATA MINING AND OFFICIAL STATISTICS

Gilbert Saporta

Chaire de Statistique Appliquée; Conservatoire National des Arts et Métiers

Abstract: Data mining is a new field at the frontiers of statistics and information technologies (database management, artificial intelligence, machine learning, etc.) which aims at discovering structures and patterns in large data sets. We examine here its definitions, tools, and how data mining could be used in official statistics.

1. What is Data Mining?

Data Mining is often presented as a revolution in information processing. Here are two definitions taken from the literature:
U.M. Fayyad : “Data Mining is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data”.
D.J. Hand : “Data Mining (consists in) the discovery of interesting, unexpected, or valuable structures in large data sets”.

The development of Data Mining is related with the availability of very large databases and the need of exploiting these bases in a new way. As J.

Friedman (1997) pointed out, the Data Base Management community has become recently interested in using DBMS for decision support. They realized that data which have been collected for management organisation such as recording transactions, might contain useful information for e.g. improving the knowledge of, and the service to customers.

One important feature is thus that Data Mining is concerned with data which were collected for another purpose: it is a secondary analysis of data bases. See for instance J.W.P.F. Kardaun and T. Alanko (1998) who speak of VLDB-NPFA, which stands for : "Very Large Data Bases with data that are collected Not Primarily For Analysis, but for the management of individual cases".

The metaphor of Data Mining means that there are treasures (or nuggets) hidden under mountains of data, which may be discovered by specific tools. Of course this is not a new idea and there has been many developments in statistical methodology which were oriented towards the discovery of patterns or models through techniques like EDA (Exploratory Data Analysis) and Multivariate Exploratory Analysis (dimension reduction methods : component, correspondence and cluster analysis among many others). J.Kettenring (former ASA president) defines statistics as "the science of learning from data".

What is new in Data Mining is :
- the availability of very large data bases, which makes obsolete classical definitions of large samples: billions of records, terabytes of data are not unusual. More and more data are collected automatically.
- the massive use of new techniques coming from the computer science community like neural networks, decision trees, induction rules.
- commercial interests in valorising existing information in order to propose individual solutions to targeted customers.
- new software packages, more user-friendly, with attractive interfaces, directed as much towards decision makers as professionals analysts, but much more expensive!

It is interesting to remark that statisticians (those who are not reluctant to this new trend…) prefer the expression Data Mining, while computer scientists use Knowledge Discovery, but it is more a matter of expressing a belonging to a community than a real difference.

2. Goals and tools of Data Mining

The purpose of Data Mining is to find structures in data. Following D. Hand (2000), there are two kinds of structures which are sought in data mining

activities: models and patterns.

Building models is a major activity of many statisticians and econometricians, especially in NSI's and it will not be necessary to elaborate too long on this. Let's say that a model is a global summary of relationships between variables, which both helps to understand phenomenons and allows predictions. Linear models, simultaneous equations are widely used. But a model is generally chosen on an a priori basis, based upon a simplifying theory. Exploration of alternative models, possibly non-linear or in a non closed mathematical form is made feasible by DM algorithms. DM appears then as a collection of tools presented usually in one package, in such a way that several techniques may be compared on the same data-set.

DM does not reduce to decision trees, or neural networks, or genetic algorithms, as some software vendors claim.

DM algorithms offers extensive possibilities of finding models relating variables together: we have mentioned neural networks, decision trees, which finds non-linear models, graphical models (or Bayesian belief networks) etc. give a valuable representation of relations between variables.

In contrast to the global description given by a model, a pattern is often defined as a characteristic structure exhibited by a few number of points: for instance a small subgroup of customers with a high commercial value, or conversely highly risked. Tools involved here are often cluster analysis, which is a well known statistical technique, but also Kohonen self-organising maps, based on neural networks.

Association rule discovery, or market basket analysis, is one of the favourite and perhaps new tools of Data Mining: its origin is in analysing purchases in supermarket: one is interested in the percent of customers who purchase simultaneously two goods, more precisely in identifying couples of goods A, B with high conditional probabilities of purchasing A, given that B is purchased. Of course the result is interesting only if P(A/notB) is low.

The problem in pattern discovery is to be sure that patterns are real and useful: the probability of finding any given pattern increases with the size of the database. It is thus necessary to develop validation rules, and in this respect processing the whole database is not the best thing to do: sampling (or splitting the base into several sets) will be better than exhaustive processing. One should test if a model or a pattern remains valid in another part of the base than the one which has been explored. Usefulness should also be tested: associations are only correlations, and not causation and it is not sure that promoting B will mechanically increase the purchase of A…

3. New mines: texts, web, symbolic data…

Besides classical data-base where data are usually presented in form of a rectangular array, new kinds of data are now present:

Symbolic data, such as fuzzy data or intervals: the data is not known precisely but belongs to an interval, with or without a probability distribution. Eurostat has financed the SODAS project (symbolic data analysis for official statistics), a consortium of 17 European research teams, and a software is available which may be used for Data Mining purposes, see Hebrail (1997) and Bock and al. (2000).

Text mining: most of the information which circulates is now digitalized as wordprocessor documents, and powerful techniques are currently available for a wide variety of applications: for example, one can categorise information from newsagencies, analyse patent portfolios, customer complaints, classify incoming e-mail according to predefined topics, group related documents based on their content, without requiring predefined classes, or assign documents to one or more user-defined categories.

Web mining is a new concept for statistical analysis of website information (text mining) or frequentation, as well as of the behaviour of the web-surfers.

4. Applications in official statistics

Of course in National Statistical Institutes there has always be some use of Exploratory Data Analysis, or of model choice algorithms, but as far as I know it seems that there are few, if not none, known applications of data mining techniques in the meaning of trying to discover new models or patterns in their databases by using the new tools described before.

It is not surprising because the main task of NSI's is data production, and analysis is often done by different institutes. Furthermore, it seems that the idea of exploring a database with the objective of finding unexpected patterns or models is not familiar to official statisticians who have to answer precise questions and make forecasts. Statistical analysis are done generally if they can be repeated in a production framework. But as far as NSI's manage large data bases on population, trade, agriculture, companies, there are certainly great potentialities in exploiting their mines of data.

Let us point out a few fields where data mining tools could be useful:

- business statistics with special mentions for innovation policy, financial health
- household equipments and savings

- health statistics, more precisely mortality and morbidity, in order to detect unexpected risk factors
- analysis of metadata information by means of text mining

The question remains of who will analyse these databases, the NSI's or other researchers? Data Mining techniques use individual records ; since confidentiality issues are crucial, it seems that NSI's are the best place, but it implies new directions of research, and likely new recruitments.

Let us mention that Eurostat is funding research projects on Data Mining and Knowledge Extraction, specially designed for the NSI's such as SODAS, KESO (Knowledge Extraction for Statistical Offices), SPIN! (Spatial Mining for Data of Public Interest). All these projects aim to provide specific software and it is very likely that within three years, specific Data Mining tools will be made available to official statisticians.

Conclusion

Data Mining is a growing discipline which originated outside statistics in the data base management community, mainly for commercial concerns. The gap is now filled and Data Mining could be considered as the branch of exploratory statistics where one tries to find unexpected and useful models and patterns, through an extensive use of classic and new algorithms. The expression "unexpected" should not be misleading: one has greater chances to discover something interesting when one is previously acquainted with the data.

Caution is of course necessary to avoid drawing wrong conclusions and statisticians, because they are familiar with uncertainty and risk, are the right people to derive validation tests.

A completely automatic process of knowledge extraction is also a misleading idea: even with very efficient software, human expertise and interventions are necessary.

Official statistics should be a field for Data Mining, giving a new life and profitability to its huge databases, but it may implies a redefinition of the missions of NSI's.

References:

J.H.Friedman (1997) Data mining and statistics: what's the connection? http://www-stat.stanford.edu/~jhf/ftp/dm-stat.ps

H.H.Bock, E.Diday (eds) (2000) Analysis of Symbolic Data. Exploratory methods for extracting statisti - cal information from complex data, Springer-Verlag

D.J.Hand (1998a) Data mining: statistics and more?, The American Statistician, 52, 112-118

D.J.Hand (1998b) Data mining-reaching beyond statistics, Research in Official Statistics, 2, 5-17

D.J.Hand (2000) Methodological issues in data mining, Compstat 2000: Proceedings in Computational Statistics, ed. J.G.Bethlehem and P.G.M. van der Heijden, Physica-Verlag, 77-85
G.Hebrail (1997) Statistical Data Analysis, Machine Learning, Data Mining, or Symbolic Data Analysis? ASMDA 1997: VIIIth International Symposium on Applied Stochastic Models and Data Analysis, Invited&Specialised Sessions Papers, 145-154.
J.W.P.F. Kardaun, T.Alanko (1998) Exploratory Data Analysis and Data Mining in the setting of National Statistical Institutes, proceedings of NTTS98 conference, http://europa.eu.int/en/comm/euro - stat/research/conferences/ntts-98
G.Saporta (1998) The Unexploited Mines of Academic and Official Statistics, in Academic and Official Statistics Co-operation, Eurostat, 11-15

Quinta
Conferenza
Nazionale
di Statistica

WORKSHOP

Basi informative integrate per l'analisi dei fenomeni sociali

coordinatore

Lorenzo Bernardi

docente dell'Università di Padova

Roma 15, 16, 17 novembre 2000
Palazzo dei Congressi
Piazzale Kennedy - EUR

ISTAT

Basi informative integrate per l'analisi dei fenomeni sociali

coordinatore

Lorenzo Bernardi

docente dell'Università di Padova

WORKSHOP

Basi informative integrate per l'analisi dei fenomeni sociali

Il Sistema Informativo Territoriale INtegrato (SIT-IN) come base per l'analisi spaziotemporale dei fenomeni sociali

S. De Francisci

Il Sistema Informativo Territoriale INtegrato (SIT-IN) come base per l'analisi spaziotemporale dei fenomeni sociali

Stefano De Francisci

Istat - Istituto Nazionale di Statistica - Via C. Balbo, 16 - 00184 Roma -
e-mail: defranci@istat.it

Sommario

Il sistema informativo SIT-IN (**S**istema **I**nformativo **T**erritoriale **IN**tegrato), realizzato in versione prototipale in ISTAT, si propone come uno strumento di base per la gestione della componente spaziotemporale dell'informazione statistica, utile in special modo come supporto nella analisi dei fenomeni sociali. Nella realizzazione del sistema ci si è posti l'obiettivo dell'integrazione dei sistemi informativi a base territoriale dell'Istituto secondo il principio della cooperazione di dati e processi in sistemi informativi complessi. Nell'ambito dei S.I. territoriali, oltre che per la progressiva tendenza a spostarsi verso logiche relazionali o ad oggetti, il problema cooperazione si manifesta specificamente al momento di trattare contemporaneamente dati descrittivi e dati geometrici, dati storici e dati attualizzati, dati proprietari e dati condivisi. D'altra parte, proprio la necessità di trattare le dimensioni "spazio" e "tempo" non più separatamente ma come un'unica dimensione complessa per gestire correttamente le trasformazioni del territorio nel tempo, ha orientato le scelte progettuali verso una logica di integrazione a più livelli, a partire da quello concettuale fino a quello fisico. Nel caso dell'ISTAT le componenti informative di base che trattano tali aspetti e che si rende necessario integrare sono costituite da: gestione spazio-temporale dei dati statistici, basi territoriali per censimenti e indagini, *address-matching* e GIS. L'integrazione avviene attraverso la realizzazione di un *middleware* complesso che costituisce l'elemento intermedio della architettura di dispiegamento *Client/Server multitier/multilayer* su cui operano i S.I. ISTAT. Il *middleware* si fonda da una parte sui principi dei "sistemi federativi ad accoppiamento leggero" e dei *datawarehouse* e dall'altra sulla definizione di un proprio livello autonomo di dati e funzioni, che costituiscono il vero valore aggiunto della logica federativa. Per giungere ad un sistema integrato si è proceduto operando contemporaneamente sui piani progettuali, applicativi, tecnologici e organizzativi, dotando ciascuno di tali piani di specifiche regole e modalità di integrazione. Per la progettazione è stata messa a punto una metodologia ad hoc; per gli aspetti applicativi si è lavorato prevalentemente sulla componente della metainfomazione per quanto riguarda i dati e sullo sviluppo per componenti per quanto riguarda le funzioni; a livello tecnologico si è puntato sui principi di *distributed object computing* con lo sviluppo di componenti distribuite, "avvicinate" tramite l'uso di un tessuto connettivo realizzato in *Java*; a livello della organizzazione generale del sistema ci si è mossi prevedendo un modello che favorisca l'interscambio controllato di servizi tra i vari utilizzatori e i sistemi collegati.

1. Introduzione

Da sempre i termini "statistica" e "territorio" sono trattati in modo strettamente interconnesso. La messa a fuoco delle informazioni su base territoriale, non a caso cominciata in modo significativo a partire dalla diffusione regolare dei censimenti della popolazione in Europa (ci si riferisce alla Francia del XIX secolo e in particolare all'opera di Quetelet e al successo del suo tentativo di

definire una scienza sociale fondata sulla statistica) ([1]), è stata anzi una delle opportunità che hanno favorito l'instaurarsi di una concezione scientifica delle "scienze umane" (secondo l'accezione di Cohen) ([2]) ed in qualche modo la stessa accettazione dei metodi dell'indagine scientifica in ambito economico e soprattutto sociale. Proprio la realizzazione dei primi censimenti ha permesso di entrare in possesso di informazioni sul territorio *fruibili*, cioè sistematiche e certificate, che, come osserva Lucio Gambi, sono lo strumento primario di cui si deve disporre quando si abbia a che fare "con carte che vogliano sospingere lo sguardo di là dagli accadimenti unici o dalle ampiture più o meno a lungo statiche delle ripartizioni amministrative, e riflettere ad esempio le strutture della società, e specialmente le condizioni di vita – alimentazione, salute, cultura -, le situazioni del popolamento, i fenomeni economici…" ([3]). Così, la definizione di una "ontologia" dello spazio (e del territorio) da una parte e quella del tempo dall'altra portano dapprima a modificarne le stesse concezioni di base, trasformando lo spazio in *prospettiva* e il tempo in *storia* ([4]) e poi a considerare l'interazione tra spazio e tempo come un fenomeno più complesso del "naturale" rapporto ortogonale, per lungo tempo accettato da scienziati e filosofi, come argutamente nota Kern riferendosi al "mollusco di Einstein" ([5]). Successivamente, un ulteriore colpo alla concezione deterministica dello spazio *organizzato* è stata data dai diversi paradigmi in base a cui nelle scienze geografiche (e – all'interno di queste – nelle varie teorie dell'analisi territoriale) si è guardato - ideologicamente - al concetto di "regione". Dalle tendenze storicistiche, fino a quelle funzionali e alle teorie sistemiche, esse hanno tutte in una qualche misura accettato la natura "vitale" del territorio, che può essere involutiva o evolutiva, rivoluzionaria o conservativa, ma è comunque sempre viva ("La struttura [territoriale] può crescere, contrarsi o restare stazionaria" ([6])). Anche la più recente "contrapposizione" tra l'approccio riduzionista e quello olista nello studio per la modellazione dei sistemi urbani ([7]), così come la visione dei sistemi territoriali come "ecosistemi" ([8]), pone comunque come fattori comuni di tutti gli approcci sistemici al territorio i concetti di complessità, evoluzione, forma.

In questo quadro è evidente che il rapporto tra spazialità e temporalità dei fenomeni sociali non può essere indagato disponendo di sole visioni di natura statico-comparata né di soli modelli dinamici di natura predittiva e simulativa, che "utilizzano" dati senza occuparsi della struttura architetturale né della gestione di tali dati. Come verrà più avanti illustrato, occorre invece costruire un sistema di riferimento che poggi su una specifica architettura dei dati, incentrata sul trattamento della dimensione spazio-temporale dell'informazione statistica. L'approccio che si propone guarda ad un percorso di integrazione articolato sul classico ciclo di vita dei sistemi informativi statistico-territoriali, (acquisizione, validazione, elaborazione e diffusione) e alle loro componenti informative di base, rappresentate da dati grezzi, microdati, macrodati, metadati e *thesauri*. Le tesi trattate nella presente nota sono dunque incentrate sul principio che l'integrazione dei sistemi informativi territoriali in ambito statistico non si identifica solo come un'esigenza organizzativa o come fatto di

([1]) *Cfr.* [9], pag. 39
([2]) *ib.* pag. 173
([3]) *Cfr.* [12], pag. 668.
([4]) "Gli infiniti e diversi aspetti del reale si classificano e ordinano in un sistema razionale, si manifestano in una forma unitaria e universale, lo spazio. Allo stesso modo si ordinano gli infiniti e diversi eventi che si succedono nel tempo. La forma o la rappresentazione secondo ragione dello spazio è la prospettiva; la forma o la rappresentazione secondo ragione del succedersi degli eventi è la storia" *Cfr.* [5], pag. 761.
([5]) "In queste circostanze, la griglia di un sistema con coordinate cartesiane è inutile per tracciare il piano del movimento: Einstein suggerì di usare un sistema di riferimento che chiamò un «mollusco di riferimento»." *Cfr.* [13], pag. 231.
([6]) *Cfr.* [17], pag. 15
([7]) Si veda ad es.: [7] e [8]
([8]) Si veda ad es.: [3], [15], [27]

"*business*" (come avviene per la maggior parte dei sistemi informativi aziendali) ma diviene un passo fondamentale per la "produzione di conoscenza".

La nota è strutturata come segue. Nella sezione 2 viene inquadrato il problema del nuovo rapporto che è possibile instaurare tra statistica e territorio: autonomia informativa, cooperazione applicativa, dimensione spazio-temporale. Nella sezione 3 viene descritto il meccanismo di integrazione adottato per il prototipo del Sistema Informativo Territoriale Integrato (SIT-IN) realizzato in ISTAT: principi della cooperazione, integrazione concettuale, architetture applicative e di dispiegamento. Infine nella sezione 4 vengono sintetizzate alcune riflessioni conclusive.

2. Un nuovo rapporto tra statistica e territorio come punto di partenza per l'integrazione dei sistemi a sfondo sociale

2.1. Necessità di una autonomia informativa della componente territoriale

Non sempre il rapporto tra produzione statistica e territorio – che come si è detto è storicamente molto stretto - viene però ad essere definito in modo chiaro. Soprattutto non sempre si tiene conto del fatto che statistica e territorio hanno vari modi di interazione. Intanto, partendo dal versante statistico e facendo riferimento al naturale percorso da compiere per realizzare efficaci indagini statistiche - articolato principalmente nelle fasi di pianificazione, preparazione, esecuzione, controllo e correzione, elaborazione, rilascio - una prima distinzione può essere individuata tra un modo di usare il territorio come base di preparazione ed esecuzione delle indagini dirette e l'altro come base di elaborazione e rilascio delle informazioni raccolte. La differenza, apparentemente banale, si rivela cruciale se si accetta l'osservazione che i due modi di interazione siano spesso inconciliabili. Quanto più la conoscenza dei fenomeni sul territorio "serve" a realizzare l'indagine (per esempio per l'operazione di campionamento areale) tanto più il territorio sarà poco "ricompensato" in termini di restituzione dell'informazione. D'altra parte la disponibilità di informazioni su base territoriale ex-ante le indagini – in genere originata dall'esistenza di giacimenti informativi "istituzionali" a larga scala territoriale – se da una parte favorisce obiettivamente la tematizzazione territoriale, dall'altra limita la libertà di costruire l'indagine stessa (i dati presenti nei giacimenti sono "rigidi" e non permettono pianificazioni sofisticate su unità di rilevazione e di analisi; in sostanza si prende quello che c'è).

Se spostiamo il punto di osservazione e ci poniamo sul versante del territorio e cerchiamo di capire se per esempio nell'analisi e pianificazione territoriale le cose siano più coerenti, non possiamo altro che verificare la stessa difficoltà. Programmi quadro, patti territoriali, contratti d'area hanno bisogno della realizzazione di efficaci attività di disegno, programmazione, attuazione, monitoraggio, valutazione degli effetti, ecc. e tali attività presuppongono dati, concetti statistici e processi di trattamento specifici e diversi gli uni dagli altri, proprio a seconda della fase di lavoro. Su tali dati e su tali processi di trattamento – in realtà – non viene invece effettuato l'adeguato investimento in termini di sistema informativo, ma essi vengono "consumati" rapidamente per le esigenze decisionali estemporanee.

La soluzione che si intravede è dunque quella di dotare il territorio di una propria autonomia e specificità informativa, costruendo dei sistemi integrati di gestione delle informazioni a base territoriale come congiunzione degli approcci spaziale-geometrico e relazionale-descrittivo; in tal modo risulta possibile sfruttare da una parte le caratteristiche e le potenzialità dei sistemi di gestione delle basi di dati (di tipo alfanumerico) – specialmente nella modellazione, nell'integrazione funzionale, nella memorizzazione, nella gestione della integrità, nell'accesso - dall'altra quelle dei

sistemi di trattamento dei dati spaziali (di tipo geometrico) – specialmente nella gestione di strati territoriali, accuratezza e fonti, in quella delle *queries* spaziali e delle specifiche operazioni collegate, quali *panning*, *zooming, buffering*, ecc.

Il percorso tracciato sinora in tale direzione non è però a sua volta ancora ben definito. Tenuto conto che, formalmente, un GIS può essere definito come un "*tool* per legare una base di dati di attributi a carte digitali"([9]), è doveroso osservare che finora, l'attenzione in ambito GIS è stata posta prevalentemente ai database che gestiscono "dati spaziali" (statici o dinamici) per applicazioni specifiche sul territorio (medicina, climatologia, ambiente, ecc.), in cui sono cioè trattati oggetti la cui posizione, forma e dimensione varia nel tempo. Vari sono gli approcci per la definizione di una "ontologia" dei sistemi a base spaziale. Tali approcci coprono vari livelli di astrazione di modellazione, andando da quello teorico, a quello concettuale, fino a quello logico([10]). In linea generale, nella definizione di una *ontologia* del territorio i principali elementi caratterizzanti sono: il tipo di ***oggetto spaziale*** (punti, linee, *regions*), il tipo di ***tempo*** (validità, transazione, esistenza), il tipo di ***variazione nello spazio-tempo*** (*moving objects*, variazioni discrete di e tra oggetti), il tipo di ***proprietà associate agli oggetti*** (spaziali o non-spaziali). La combinazione delle modalità dei vari aspetti fornisce un quadro complesso della informazione definibile nei sistemi a base territoriale, la cui corretta organizzazione risulta centrale nel tentativo di costruzione di basi di dati evolute.

In generale, quindi, da una parte l'approccio sistemico alla modellazione territorio non ha tenuto nel giusto conto le prospettive offerte dalle basi informative e se lo ha fatto non ha distinto tra la componente specifica dello spazio e quella del tempo, né tantomeno ha posto attenzione alla descrizione vera e propria degli *oggetti spaziotemporali*; dall'altra, una volta arrivati a tale distinzione tramite la costruzione di STDBMS ([11]), non è stata invece posta sufficiente attenzione alla componente *statistica* dell'informazione (sintetizzabile nella coesistenza di microdati, macrodati e metadati e delle specifiche funzioni di trattamento).

2.2. La cooperazione tra sistemi informativi eterogenei come soluzione per l'integrazione delle basi informative

In generale, l'obiettivo della cooperazione tra sistemi è quello di mantenere l'indipendenza di ogni singolo "dominio" applicativo e di mettere in condivisione "domini" diversi.

Nel caso dell'ISTAT la cooperazione applicativa deve essere imperniata su un fulcro costituito dai sistemi informativi che trattano l'informazione statistica e sui sistemi che ruotano attorno ad esso (sistemi di dati *reference*, *thesauri* statistici, dati a base territoriale, basi di dati e strati software comuni, ecc.); questi ultimi, cooperando col sistema statistico come sistemi satellite, mettono a disposizione un forte valore aggiunto, rappresentato sia dal corredo di metainformazioni di riferimento sia da alti livelli di riusabilità e generalizzazione applicativa.

Un altro elemento importante è la navigazione trasversale attraverso informazioni appartenenti ai vari sistemi. In questa prospettiva, il sistema informativo del territorio può costituire per sua natura uno degli elementi portanti nella cooperazione tra i sistemi informativi statistici, perché può

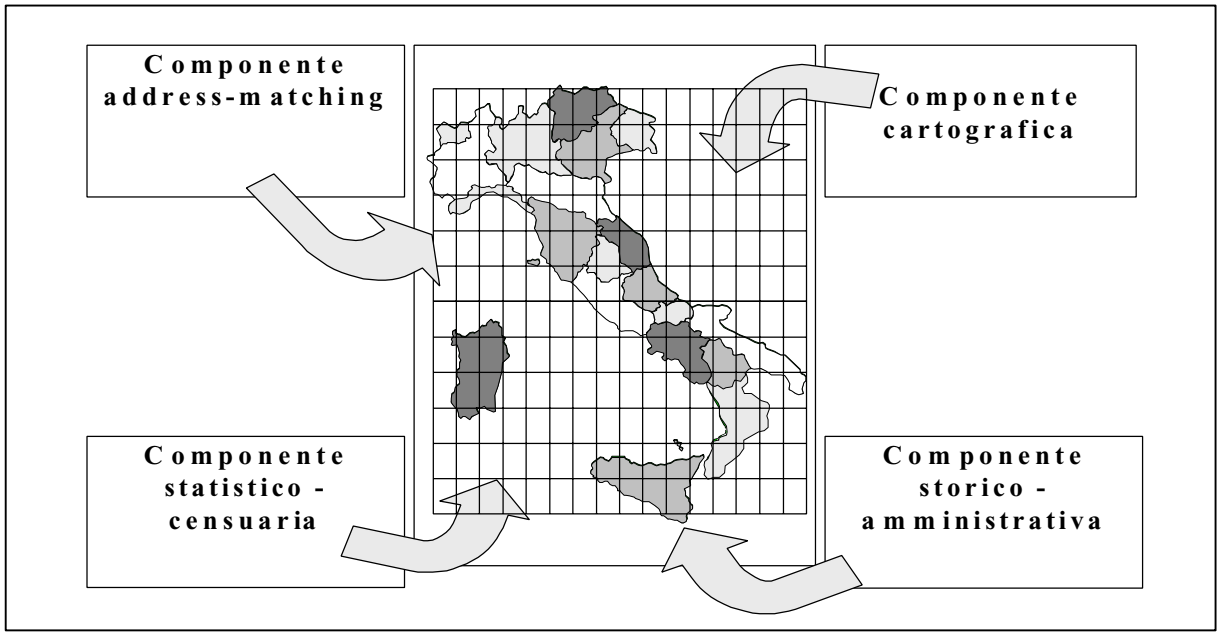
([9]) Definizione riportata in [16]
([10]) Si può vedere, rispettivamente, per gli aspetti teorici [29]; per il livello concettuale [24]; per il livello logico [23]
([11]) STDBMS sta per Spatiotemporal Database Management System

garantire il diretto e corretto uso delle dimensioni spaziale e temporale nel trattamento delle informazioni prodotte.

Proprio l'esperienza della cooperazione tra varie componenti che trattano dati a base territoriale realizzata nel sistema informativo territoriale integrato (SIT-IN) costituisce un esempio di questa strategia di cooperazione attraverso la dimensionespazio-temporale. Il sistema realizzato è stato infatti il tramite immediato attraverso cui è stato possibile integrare un sistema informativo di diffusione di dati statistici come quello rappresentato dalle Banche Dati Territoriali, che raccoglie dati statistici provenienti dagli ultimi tre censimenti e da altre indagini, aggregati per comune e interrogabili ora in maniera interattiva, dinamica e flessibile.

Il sistema informativo territoriale integrato poggia sulla interazione di sistemi di base che coprono molteplici aree informative (v. fig. 1):

Figura 1 - Il quadro dei sistemi a base territoriali dell'ISTAT



- la componente per la gestione delle trasformazioni del territorio nel tempo, dal punto di vista amministrativo e statistico (il sistema SISTAT),
- le componenti per la gestione delle basi territoriali dei censimenti e della cartografia digitale (i sistemi CENSUS 2000 e GISTAT),
- la componente diaddress-matching (il sistema SISTER),
- la componente statistica costituita da dati e metadati (nel prototipo sono stati considerati database relazionali esterni e il sistema Banche Dati Territoriali);

tutti questi sistemi - a loro volta disponibili in versioni prototipali o in via di completamento o di reingegnerizzazione - vengono federati attraverso la componenti dati, cooperano applicativamente tra loro e - oltre ad avere interfacce proprie - si prevede possano essere presentati agli utenti anche tramite un'interfaccia unica.

2.3. Una visione unitaria del problema "spazio-temporale" dell'informazione statistica

Elemento centrale dell'integrazione è costituito proprio dal meccanismo di trattamento di "spazio" e "tempo". Nei database "di produzione statistica" la corretta definizione del riferimento territoriale nel tempo (o dualmente la corretta definizione temporale nelle dinamiche spaziali) è sempre uno dei primi passi nella validazione dei dati, mentre nelle strutture informative "di diffusione" (una delle cui strutture di riferimento oggi più frequente è rappresentata dall'organizzazione dei dati secondo un approccio basato su *datawarehouse* e database *multidimensionali*) le due dimensioni presenti costantemente sono proprio lo spazio e il tempo.

Le esperienze secondo gli approcci tradizionali hanno fino ad ora considerato le due dimensioni come separate rispetto al dato di riferimento. Una visione unitaria dello spazio e del tempo (attraverso la introduzione di una dimensionespaziotemporale) rappresenta invece per i sistemi informativi statistici un elemento comune attorno al quale si possono riferire dati appartenenti a sistemi eterogenei ed è un presupposto attraverso il quale è possibile realizzare la cooperazione tra diverse realtà. Di seguito verranno esaminati alcuni aspetti di questa soluzione.

3. Il meccanismo dell'integrazione adottato in SIT-IN

3.1. Principi della cooperazione applicativa adottati in SIT-IN

Oltre ai passi tradizionali che possono essere individuati per la integrazione dei sistemi, l'integrazione in ambito statistico dei sistemi a base territoriale è caratterizzata da specifiche fasi. Per quanto riguarda le problematiche classiche dell'integrazione, l'attenzione è stata rivolta per lo più al problema dell'integrazione dei dati ([12]) o alla soluzione di singoli contesti applicativi ([13]) mentre, in una trattazione del problema nel suo insieme, quale è quella alla base della realizzazione di SIT-IN ([14]), l'integrazione di sistemi informativi a base territoriale può essere vista come un processo composito che riguarda più "oggetti" (regole, dati, applicazioni, tecnologie) e più "livelli" (semantica, schema, formato, fisico); i vari livelli e oggetti sono interdipendenti e l'integrazione risulta caratterizzata da vincoli intra-livello e inter-livello. In particolare, mentre per quanto riguarda regole e concetti (livelli "semantico" e "di schema") l'integrazione è rigida e garantisce la completa disambiguazione dei sistemi, per quanto riguarda applicazioni e tecnologie (livelli "di formato" e "fisico") l'integrazione è prevalentemente basata suinteroperabilità e cooperazione.

I passi del percorso metodologico adottato per l'integrazione in SIT-IN si orientano in sostanza verso un duplice obiettivo: (a) integrazione a livello concettuale dei sistemi; (b) costruzione di un sistema informativo integrato. Il primo permette una visione unitaria dell'informazione, e prevede come punti cardine l'adozione di un livello comune di rappresentazione concettuale dei dati spaziali e descrittivi nel sistema informativo territoriale (entità, relazioni, vincoli e regole di integrità) e l'organizzazione dei concettispazio-temporali in specifiche strutture concettuali. Il secondo invece avviene attraverso tre meccanismi: graduazione del livello di incapsulamento dei sistemi base, massiccio uso di metodi diinterconnessione e intermediazione applicativa, unificazione dei criteri di interfacciamento da e verso l'esterno.

([12]) Per una trattazione approfondita dei vari approcci, si veda [1], [6], [14], [19], [21]
([13]) Si veda [2]
([14]) La trattazione della metodologia di progettazione di SIT-IN è riportata approfonditamente in [4]

Per il raggiungimento dei due obiettivi si è adottato il paradigma della cooperazione applicativa sia per i dati sia per le funzioni. Per i dati si è trattato di lavorare per una integrazione concettuale fondata sulla dimensione ***spaziotemporale*** come riferimento unico per tutti i sistemi, sulla realizzazione di un sistema federato ad accoppiamento "*leggero*" per metadati e dati comuni e sulla creazione di un sistema a replicazione di dati (*datawarehouse*) per la diffusione. Ai fini della cooperazione a livello di processi le soluzioni adottate vertono su architettura distribuita di tipo *multi-tier*, sviluppo per componenti, *middleware* applicativo di intermediazione tra funzioni/ambienti, interfacce WEB.

L'approccio di integrazione concettuale permette di trattare contestualmente nella modellazione sia gli aspetti geometrici sia quelli alfanumerici, ambedue variabili nel tempo, e di definire tra tali classi di concetti i necessari vincoli di integrità e di cardinalità. Partendo da uno schema astratto e dando corso ad una progressiva serie di raffinamenti concettuali si perviene ad un modello concettuale di dettaglio in cui tutti i concetti della realtà sono adeguatamente rappresentati nello schema e non ulteriormente raffinabili. Questo schema è il punto di partenza per la successiva fase di progettazione logica. Solo a questo punto si è obbligati a procedere lungo due strade separate: da una parte i concetti di natura geometrica (gli oggetti spaziali), trattati dagli strumenti GIS, dall'altra i concetti descrittivi, trattati in ambiente relazionale.

3.2 Integrazione a livello concettuale.

3.2.1. L'approccio classico nella modellazione dei dati geografici

In generale, mentre per la componente alfanumerica dell'informazione ci si avvale di un passo di modellazione concettuale, per i dati geometrici ci si limita tutt'al più – prima di passare all'implementazione - a definire un passo di modellazione logica.

Rimanendo in ambito di STDBMS e GIS, riprendiamo la definizione di "oggetto spaziotemporale", come riportata in letteratura ([15]):

"Un oggetto spazio-temporale O (dichiarato dal suo identificativo O_id) è un oggetto spaziale evolvente nel tempo, cioè, la sua evoluzione (o storia) è rappresentata da un insieme di triple (O_id, s_i, t_i), dove , s_i è la localizzazione dell'oggetto O_id all'istante t_i (s_i e t_i sono chiamati rispettivamente 'space-stamp' e 'time-stamp')".

A partire da tale definizione, è possibile classificare come sistemi spaziotemporali un complesso di diversi contesti, tra i quali si può per brevità citare:

- Sistemi che gestiscono oggetti spaziali (punti, linee, regions) che si evolvono nel tempo (*time-evolving spatial objects*)
- Sistemi che gestiscono gli intervalli di validità temporale degli oggetti nel tempo (*valid time systems*) o sistemi che gestiscono gli istanti di trasformazione degli oggetti nel tempo (*transaction time systems*)
- Sistemi che gestiscono oggetti "fermi" nello spazio (*still objects systems*) o sistemi che gestiscono oggetti in movimento (*moving objects systems*)
- Sistemi a cardinalità statica o dinamica nel tempo

([15]) cfr. [22]

Tutti i sistemi sopra citati si basano sullo stesso concetto di oggetto spazio-temporale, e tutti possono essere considerati come sistemi di tipo GIS. Pure essi hanno problematiche applicative e tecnologiche totalmente distinte che ne modificano campi di applicazione, potenzialità e usi.

3.2.2. L'approccio utilizzato per SIT-IN

In SIT-IN, mettendo a frutto le impostazioni teoriche di tali problematiche classiche, si è proceduto intervenendo proprio sulla ontologia del concetto di spazio-tempo con riferimento ai dati statistici. A livello concettuale le due caratteristiche vengono intanto fuse in una sola dimensione denominata dimensione "spaziotemporale". La dimensione spaziotemporale che vede l'evoluzione dello spazio nel tempo, assicura l'analisi del dato statistico sul territorio anche in funzione delle modifiche che esso subisce. Questa componente è importante quando si costruiscono "serie storico-territoriali" attraverso le quali si può effettuare l'analisi di un fenomeno statistico nel tempo su una stessa porzione di territorio.

Infatti la suddivisione amministrativa del territorio varia nel tempo e ad una stessa "unità amministrativa" comunemente intesa (regione, provincia, comune, ecc.) corrisponde – nel tempo - un elemento con "caratteristiche" differenti; questa complessità si riflette sul dato statistico ad esso riferito, in termini di validità e di interpretazione. Lo studio di una unica dimensione spaziotemporale ha permesso di gestire tale complessità, definendo ogni singola unità amministrativa (ad esempio il comune) come un concetto di più alto livello generato dalla fusione di concetti definiti ad un livello più elementare: i "codici", le "forme", le "gerarchie" e le inter-relazioni del territorio nel tempo.

Ad esempio, al concetto di "comune" sono legate tre distinte componenti informative:
1) la componente relativa ai codici identificativi ISTAT e alle denominazioni del comune
2) la componente relativa alle sue caratteristiche cartografiche (linee e poligoni)
3) la componente relativa alle sue connessioni nel tempo con le unità amministrative adiacenti nella gerarchia, sia a livello inferiore (sezioni di censimento) sia superiore (province), ma anche alle inter-relazioni di tipo non inclusivo con altre gerarchie e coperture.

Poiché nel tempo le tre componenti elementari possono variare indipendentemente l'una dall'altra, i dati statistici rappresentati su una determinata area territoriale variano, oltre che per le naturali dinamiche del fenomeno, anche per effetto della modifica dell'area di riferimento rispetto ad una delle tre componenti. La dimensione spaziotemporale tiene conto di queste modifiche nel tempo ed è in grado di segnalarle in relazione all'andamento del fenomeno.

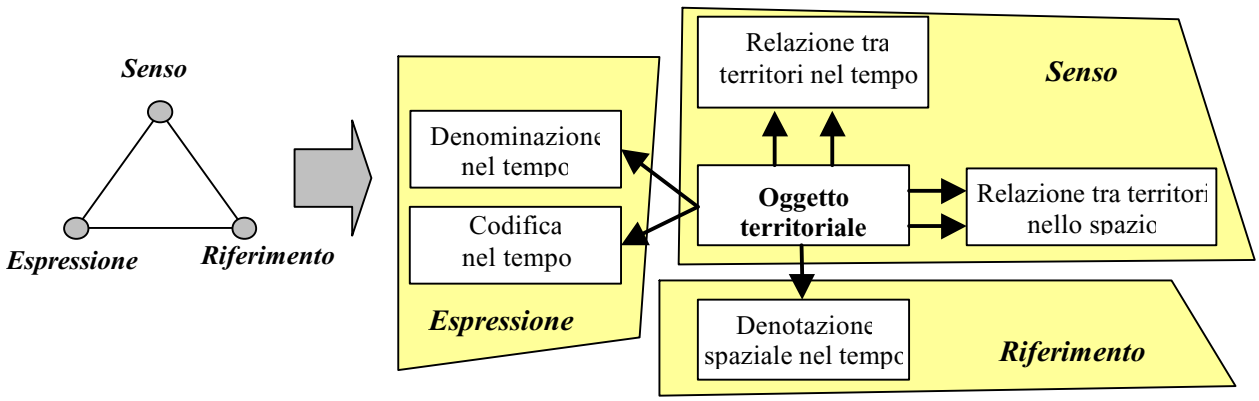
La descrizione dello schema concettuale viene dapprima proposta mediante la collocazione dei concetti spazio-temporali in una prospettiva di natura semiotica. Il principio è quello di introdurre tra i concetti dello schema ER riferiti alla componente spazio-temporale una distinzione a livello semasiologico e di modellare le diverse componenti in base al rispettivo ruolo concettuale che ciascuna di esse ricopre a livello semiotico e che, in termini non formali ([16]), può essere classificato in: espressione, senso (o significato) e riferimento reale. Nella figura seguente viene illustrato il

([16]) Nella nota non viene fatto esplicito riferimento ad alcuna linea di pensiero in ambito semiotico. Di conseguenza anche i termini utilizzati (senso, espressione, significato, denotazione, riferimento, ecc.) che nella letteratura riassumono specifiche posizioni teoriche, devono essere interpretati nella loro rispettiva accezione comune e convenzionale. Volendo comunque dare un riferimento teorico, può essere assunta l'impostazione riportata in [26] pp. 12 *sgg.*, in cui si dà conto delle principali distinzioni storiche tra i termini Espressione, Contenuto e Referente, pur mantenendo per essi anche un contesto ed una accezione non formale.

meccanismo secondo cui avviene l'organizzazione in chiave semiotica dei concetti di natura spazio-temporale nello schema concettuale.

La strutturazione introdotta permette di attivare nei concetti dello schema ER un processo di significazione, in cui cioè siano riconoscibili sia il piano dei contenuti che pertinentizzano le entità territoriali (rappresentato dagli oggetti "in sé", dalle relative marche semantiche e dalle "relazioni" spazio-temporali individuabili tra gli oggetti territoriali stessi) sia il piano delle espressioni (denominazioni e codifiche che nel tempo possono essere associate ad una stessa unità di territorio), sia il piano dei riferimenti reali (in sostanza le porzioni di territorio occupate da una entità territoriale nello spazio e nel tempo).

Figura 2 : Organizzazione in chiave semiotica dei concetti spaziotemporali

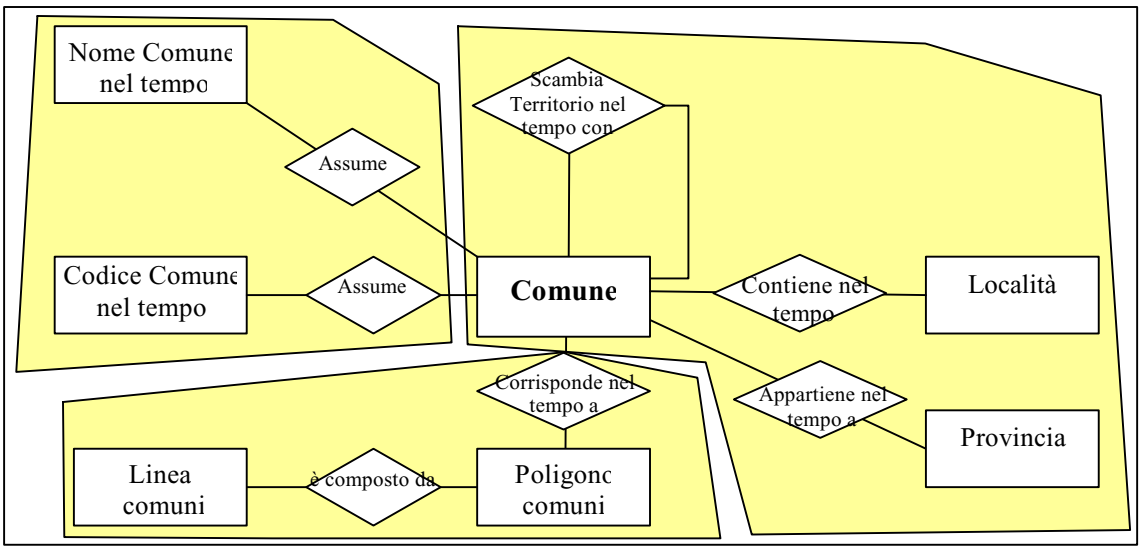


Il partizionamento così introdotto dona al modello concettuale la potenza espressiva necessaria per la rappresentazione di un mondo reale complesso, dove gli oggetti territoriali si evolvono nel tempo mutando riferimenti espressivi, relazioni reciproche spaziali e temporali, e con esse la loro forma, ma mantenendo una identità lungo l'arco del processo. Nel modello proposto, ogni oggetto territoriale O viene ad essere definito dalla terna complessa $O_i=(E_i,S_i,R_i)$, composta da un insieme di espressioni E_i, di marche semantiche S_i, e di riferimenti spaziali R_i, in cui a loro volta $S_i=(Id\_O_i,R^t_{ij},R^s_{ij})$, $E_i=(D^t_i,C^t_i)$ e $R_i=(P^t_i, L^t_i)$, dove:

Id_O_i : identificativo oggetto O_i	D^t_i : denominazioni dell'oggetto O_i nel tempo *t*
R^t_{ij} : relazione nel tempo *t* tra O_i e O_j	C^t_i : codici associati all'oggetto O_i nel tempo *t*
R^s_{ij} : relazione nello spazio *s* tra O_i e O_j	P^t_i : poligoni denotanti l'oggetto O_i nel tempo *t*
	L^t_i : linee riferite all'oggetto O_i nel tempo *t*

Attualizzando nel prototipo di SIT-IN la struttura sopra descritta, si è ottenuto uno schema concettuale in cui vengono definite per ogni oggetto territoriale le relazioni di contenimento e intersezione spaziale, le trasformazioni storiche, le varie denominazioni e i vari codici di natura amministrativa, le geometrie (poligoni, linee, punti) correnti e passate. In figura 3 viene riportato un frammento dello schema ER di dettaglio relativo all'entità "Comune".

Figura 3: Schema concettuale per l'entità Comune



Dalla figura si può notare che non solo sono stati trattati contestualmente sia i concetti "geometrici" sia quelli alfanumerici, ma sono state definite anche le relazioni reciproche tra tali classi di concetti. Dando corso ad una progressiva serie di raffinamenti concettuali – cioè fino a che tutti i concetti della realtà sono adeguatamente ed esaurientemente rappresentati nello schema – si perviene al modello concettuale di dettaglio, punto di partenza per la successiva fase di progettazione logica. Solo a questo punto si è obbligati a procedere lungo due strade separate: da una parte i concetti di natura geometrica (gli oggetti spaziali), trattati dagli strumenti GIS, dall'altra i concetti descrittivi, trattati in ambiente relazionale.

Lo schema presentato copre le problematiche di tipo "operazionale" del sistema informativo (propri dei sistemi anche detti OLTP ([17])), quelle cioè che permettono di "amministrare" ed "elaborare" i dati di base. Successivamente, al momento della aggregazione e diffusione dei dati, gli stessi concetti spaziotemporali, finora trattati in una "forma piatta", vengono strutturati in una forma che favorisce la navigazione e il ritrovamento, tipicamente nei sistemi di tipo "informazionale", o OLAP. Ciò avviene operando una trasformazione del "ruolo" che i concetti giocano nello schema dati, nella fattispecie attraverso la individuazione di "concetti-fatto" e "concetti-dimensione". Tale distinzione è alla base della costruzione di approcci al disegno delle basi di dati relazionali – denominati *star-schema* e *snowflake-schema* – che favoriscono la visione in chiave multidimensionale dei dati stessi. Verso questa trasformazione si è proceduto all'atto di integrare i dati statistici delle serie territoriali (Banche Dati Territoriali) in SIT-IN.

3.3. Integrazione e cooperazione attraverso la *federazione* dei contenuti informativi

Una possibile metodologia di cooperazione tra sistemi informativi passa attraverso la federazione della componente metadati dei vari sistemi. Infatti ognuno dei sistemi componenti mette a fattore comune la propria componente metadati come vista federata di dati locali che afferiscono al

([17]) *OLTP* sta per *On-Line Transaction Processing*, mentre *OLAP* sta per *On-Line Analytical Processing*.

database comune di dati e metadati. Questa metodologia ha consentito in fase di sviluppo di procedere ad una navigazione comune attraverso i metadati di ognuno dei sistemi componenti.

L' elemento spaziotemporale è la componente centrale di questa federazione perché permette di selezionare un elemento territoriale nel tempo e connettere ad esso gli altri contenuti informativi sia di natura cartografica sia di natura statistica.

3.4. Integrazione e cooperazione attraverso le architetture tecnologiche

Un altro elemento importante nel processo di cooperazione tra i sistemi è la componente tecnologica (v. fig. 4).

Attualmente le architetture di dispiegamento offrono vari livelli di cooperazione tecnologica, sia con il colloquio diretto tra le varie componenti, sia attraverso strati intermedi di software. Nel prototipo di SIT-IN sono state utilizzati per il *back-end* l' ambiente *Oracle*, mentre il *front-end* è stato realizzato attraverso la cooperazione tra componenti *Java* e *SAS*.

L' alto grado di complessità del contesto applicativo non ha permesso di realizzare la semplice e classica architettura *client-server*. Il prototipo di SIT-IN è infatti caratterizzato dall' architettura *multitier-multilayer* descritta in figura e di cui si dà di seguito una sintetica descrizione ([18]):

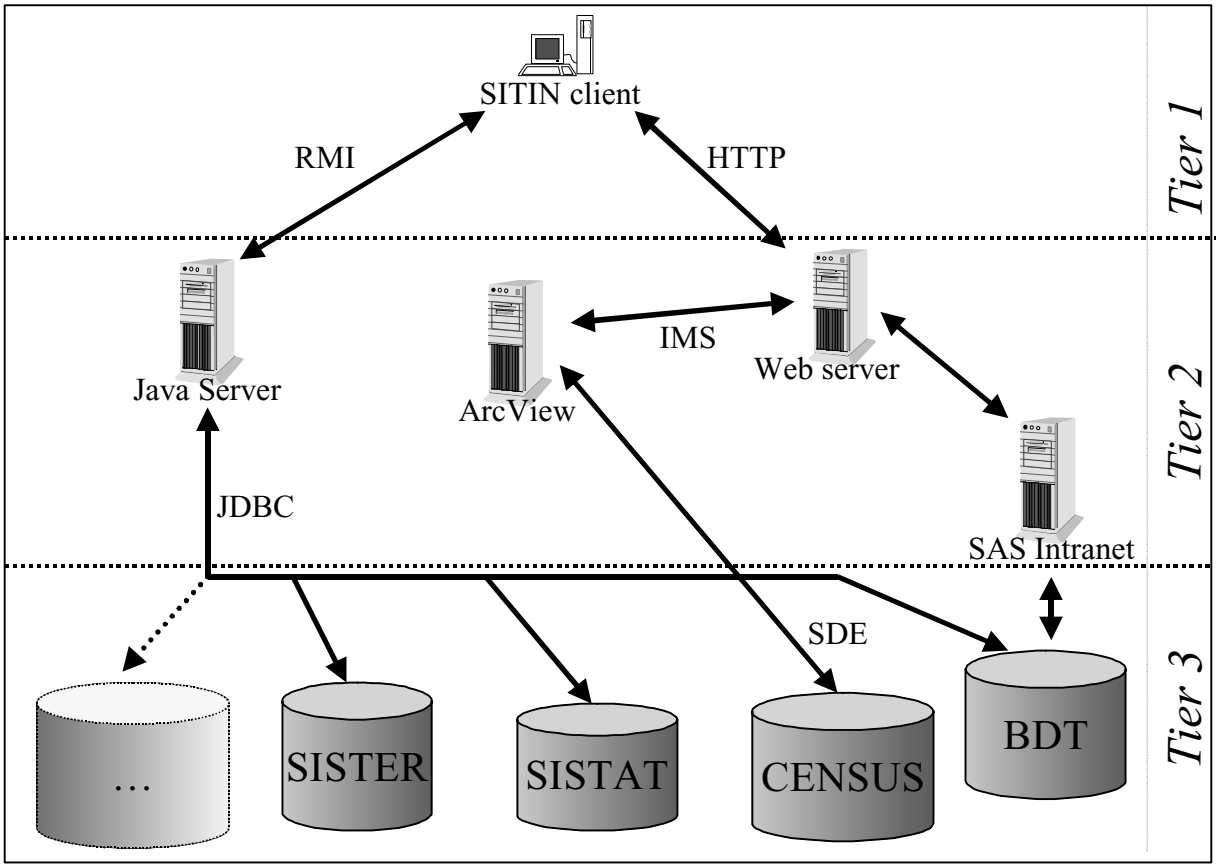


Figura 4 – L' architettura di dispiegamento di SIT-IN

([18]) Descrizioni più approfondite degli aspetti tecnologici di SIT-IN sono reperibili in [11] e [20]

***Tier* 3.** Il livello dei dati è costituito da alcune basi di dati relazionali. La base di dati cartografica è gestita tramite un motore di dati spaziali, che permette la gestione dei dati stessi tramite un modello logico-fisico relazionale. Per restare fedeli ai principi suesposti è stato necessario fare un massiccio uso di "mediatori" ([19]) per "incartare" e interconnettere i vari sottosistemi. L'uso del linguaggio *Java* e del meccanismo di connessione ai database ([20]) ha permesso di realizzare facilmente componenti di mediazione per ciascun membro della federazione.

***Tier* 2.** Il livello applicativo è il più complesso dei tre ed è composto da vari sistemi differenziati. Da una parte i *server* applicativi intermedi connettono i *client* alle sorgenti di dati alfanumerici, in modo tale da non condizionare l'interfaccia nell'accesso ai dati descrittivi, mentre dall'altra un diverso percorso di collegamento è utilizzato per il recupero delle rappresentazioni spaziali degli oggetti.

***Tier* 1.** Il *client* di interfaccia utente è uno applet *Java*. Questo comporta per l'utente il dover installare un modulo aggiuntivo per estendere il navigatore Web utilizzato e questa è l'unica componente software da installare a livello *client* nel sistema. Il livello *client* può essere considerato a sua volta come un ulteriore fattore di integrazione, visto che è proprio a tale livello che si "incontrano" e si integrano dati descrittivi e dati spaziali.

Una particolarità di questo approccio è che, grazie all'indipendenza degli strati, il numero e il tipo delle istanze di base di dati accedute dal sistema non è predeterminato alla partenza del sistema. Questa caratteristica permette agli utenti di collegare dinamicamente dati propri al sistema, al fine di controllarne la coerenza amministrativa e/o di effettuare analisi spaziali su di essi. Fornendo al sistema le necessarie informazioni di connessione, è possibile integrare dinamicamente una tabella di una base di dati esterna al sistema che contenga un insieme di dati statistici che si riferiscono a entità territoriali. L'integrazione è resa possibile dall'esistenza di un'informazione specifica del dominio applicativo: il codice territoriale. Si tratta di un codice standard che viene utilizzato dall'Istituto per identificare le entità amministrative territoriali. Questa caratteristica, insieme col fatto che anche i dati spaziali sono memorizzati in tabelle relazionali, permette di recuperare la rappresentazione geografica delle entità territoriali e la conseguente visualizzazione della mappa tematica. Proprio la possibile modifica nel tempo di tali codici è uno dei fattori che hanno reso necessario superare la separazione concettuale tra la dimensione dello spazio e quella del tempo, secondo l'approccio descritto in precedenza. In tal modo, agganciando i dati esterni tramite una sola coppia di attributi, codice identificativo e tempo, vengono ad essere ereditate tutte le altre proprietà pertinenti che permettono di definire non solo l'*identità* ma l'*essenza* stessa dei territori associati ai codici validi alla data di riferimento (denominazione, forma, relazioni con il resto del territorio, ecc.).

Inoltre, è interessante notare che il processo di integrazione dinamica è realizzato da un sistema di "meta-interrogazioni" ([21]) che è parte del sistema di gestione di metadati di SIT-IN: un insieme di tabelle che contengono dati sugli utenti, sulle tabelle esterne e i relativi valori collegati e, soprattutto, dati che descrivono la gerarchia amministrativa del territorio italiano. Per ciascuna entità territoriale astratta (nazione, regione, provincia, ecc.) è stata memorizzata una istanza nella meta-base di dati. Tale istanza comprende descrizioni testuali, informazioni sui legami tra entità e,

([19]) Si veda [10]
([20]) *Java Data Base Connectivity (JDBC)* : una suite che mette a disposizione funzioni predefinite per l'accesso generalizzato a basi di dati relazionali.
([21]) Il tema è trattato, ad es., in [18] e [25]

come elemento maggiormente caratterizzante, alcuni campi contenenti le interrogazioni in formato parametrico che permettono di eseguire le operazioni di interrogazione. L'approccio guidato dai metadati è molto utile non solo perché ha permesso di realizzare facilmente il processo di integrazione dinamica, ma anche perché favorisce l'aperturaarchitetturale del sistema. Infatti, una delle linee di sviluppo correnti del progetto è l'estensione del sistema a supportare differenti gerarchie territoriali.

Ma la caratteristica forse più originale di questa architettura è l'esistenza dei due distinti percorsi di comunicazione: il "percorso geografico" e il "percorso relazionale". L'unico punto di connessione tra i due percorsi è – come detto - a livello del*client*; questo a dispetto del fatto che i dati spaziali e i dati alfanumerici sono memorizzati nella stessa base di dati. La coerenza dei dati, indipendentemente dalla progettazione delle connessioni, è assicurata dall'integrazione a livello della base dei dati. In questo modo, dal punto di vista applicativo, un percorso di connessione permette all'utente di selezionare i dati di interesse mantenendo impliciti i riferimenti alle corrispondenti rappresentazioni geometriche; l'altro percorso, che deve essere esplicitamente attivato dall'utente, permette di accedere al motore spaziale, di visualizzare la rappresentazione geografica degli oggetti territoriali e di effettuare operazioni spaziali su di essi. Poiché le operazioni spaziali sono sostanzialmente più complesse dal punto di vista dell'elaborazione, la completa separazione applicativa porta due vantaggi rilevanti:*i)* è possibile una separazione netta dei tempi di elaborazione; *ii)* è possibile sostituire l'intero sistema di elaborazione spaziale senza toccare i sistemi di selezione delle aree territoriali e gran parte dell'interfaccia.

4. Conclusioni

L'aspetto più interessante dell'esperienza che ha portato alla realizzazione di SIT-IN è costituito dal fatto che per arrivare alla "migliore" integrazione si è dovuto intervenire su problematiche che riguardavano vari livelli di progetto, a partire da quelli di impostazione teorica (le "ontologie" alla base dei sistemi da integrare, in particolare relative ad aspettispaziotemporali e ad aspetti statistici) per passare a quelle di analisi e disegno, fino alle soluzioni relative alle componenti applicative e tecnologiche. Per arrivare alla integrazione finale del sistema, in buona sostanza, è stato necessario integrare tra loro alcuni "metodi" di integrazione, così da trattare i possibili diversi meccanismi di integrazione definiti ai vari livelli come un unico percorso complesso.

In chiave teorica è interessante notare come da una visione separata di database statistici da una parte e di sistemi di basi di dati spaziotemporali, in SIT-IN si è pervenuti ad una visione più articolata in cui – attraverso i meccanismi di integrazione descritti nella nota – i due mondi sono visti in maniera unitaria.

In chiave concettuale, elementi salienti sono, da una parte, la definizione di una dimensione spaziotemporale, che permette di disporre e trattare informazioni territoriali non solo relative alle identità dei vari oggetti spaziali, ma soprattutto alle loro "essenze"; dall'altra lamodellazione unitaria sia per la componente geografica sia per la componente descrittiva. Questo fattore permetterà, tra l'altro, di utilizzare ed integrare gli standard dimodellazione (ISO/IEC e CEN/TC) dei metadati per entrambe le aree di dati.

In chiave applicativa ci si è ampiamente soffermati sulle potenzialità offerte dai principi di cooperazione applicativa. Per i dati, da una parte, l'approccio federativo permette di mantenere basi di dati distinte ma armonizzate, allineate e convergenti verso una omogeneità*de facto* dei dati, mentre, dall'altra, l'approccio *datawarehouse* permette non solo di disaccoppiare il momento della amministrazione interna dei dati da quella dell'uso da parte degli utenti ma anche di sfruttare

appieno le caratteristiche intrinseche dei modelli dei dati a supporto delle varie componenti, nel presupposto che ogni componente del sistema abbia bisogno di specifiche "forme" nella modellazione delle informazioni. Per le funzioni, oltre a tutte le considerazioni già riportate nel testo, resta da sottolineare come la logica dello sviluppo per componenti favorisca il progressivo ampliamento del sistema nelle sue evoluzioni future e lo ponga come base di integrazione dei sistemi che trattano dati statistici.

In chiave tecnologica, infine, è risultato centrale e determinante l'uso di strumenti ed ambienti avanzati, rappresentati da WEB, *Java* e *driver* JDBC, la specializzazione dello strato di *middleware* applicativo per il trattamento di dati estremamente eterogenei (dati geometrici, dati statistici, metadati, dati "*legacy*") nonché l'uso generalizzato del database relazionale per il *back-end* di tutti i dati.

RIFERIMENTI BIBLIOGRAFICI

[1] P.H. Aiken, *Reverse Engineering of data*, IBM Systems Journal, 37 2, IBM 1998

[2] P. Aiken, O.K. Ngwenyama, L. Broome, *Reverse Engineering New Systems for Smooth Implementation*, IEEE software, 1999

[3] P.M. Allen, *Evolution, modelling and design in a complex world* – Environment and Planning B, 9, 1982

[4] A. Ambrosetti, S. De Francisci, M. Paolucci, G. Sindoni, *Integrazione di sistemi informativi territoriali: un approccio metodologico*, Atti XXXVII Congresso Annuale AICA, Abano Terme, 1999

[5] G.C. Argan, M. Fagiolo, *Storia d'Italia*, Vol. 1, *I Caratteri originali, Premessa all'arte italiana*, cap. III. *I problemi della città. La struttura chiusa*, Einaudi, Milano, 1972

[6] C. Batini, M. Lenzerini, S. B. Navathe, *A Comparative Analysis of Methodologies for Database Schema Integration*, ACM Computing Surveys 18 4, 1986

[7] C.S.Bertuglia, *La città come sistema*, in C.S. Bertuglia, A. La Bella (a cura), *I sistemi urbani*, Vol. I, *Le teorie. Il sistema e le reti*, Franco Angeli, Milano, 1991, pp. 301-309

[8] C. S. Bertuglia, G. A. Rabino, *L'uso dei modelli matematici nelle valutazioni dei fenomeni in campo urbano: presupposti concettuali e problematiche operative*, 1992

[9] I.B.Cohen, *Scienze della natura e scienze sociali*, Laterza, Bari, 1993

[10] T. Critchlow, M. Ganesh, R Musick, *Metadata Based Mediator Generation*, in Proceedings of the Third IFCIS Conference on Cooperative Information Systems (CoopIS'98), 1998, pp. 168-176

[11] S. De Francisci, M. Paolucci, G. Sindoni, L. Tininini, *Cooperazione applicativa nell'ambito dei S.I.T.: il Sistema Informativo Territoriale Integrato (SIT-IN) dell'ISTAT*, in Informazione Geografica: innovazione e formazione, Atti della 4ª Conferenza Nazionale ASITA, Genova, Italia, 3-6 Ottobre 2000

[12] L. Gambi, *Storia d'Italia*, Vol. 6, *Atlante, Parte quinta. Immagini statistiche dell'Italia unita*, Einaudi, Milano, 1976

[13] S. Kern, *Il tempo e lo spazio*, Il Mulino, Milano, 1998

[14] B. Kramer, M. Papazoglou, H.W. Schmidt, *Information Systems –Interoperability*, eds., Research studies press ltd., 1997

[15] S. T. Lombardo, *Recenti sviluppi nella modellistica urbana*, in C.S. Bertuglia, A. La Bella (a cura), *I sistemi urbani*, Vol. II, *I modelli, gli strumenti di supporto alle decisioni, i processi di governo*, Franco Angeli, Milano, 1991, pp. 642-708

[16] Brian Menneke, "*Understanding the role of Geographic Information Technologies in Business: Application and Research Directions*", JGIDA, vol. 1, no.1 pp. 44-68, 1997

[17] R. Mainardi, *Geografia regionale*, La Nuova Italia Scientifica, 1994

[18] F. Neven, J Van den Bussche, D. Van Gucht, G. Vossen, *Typed Query Languages for Databases Containing Queries*, Information Systems 24(7), 1999, pp. 569-595

[19] J. Noll, W. Scacchi, *Integrating Diverse Information Repositories: A Distributed Hypertext Approach*, Computer, IEEE 1986

[20] M. Paolucci, G. Sindoni, S. De Francisci, L. Tininini, *SIT-IN on Heterogeneous Data with Java, HTTP and Relations*, NET.OBJECT Days 2000, Erfurt 9-12 Ottobre 2000, pp. 627-630

[21] A. Silberschatz, M. Stonebraker, J. Ullman, *Database Research: Achievements and Opportunities Into the 21st Century*, Eds., NSF Workshop on the Future of Database Systems Research, 1995

[22] Y. Theodoridis, T. Sellis, A.N. Papadopoulos, Y. Manolopoulos, *Specifications for Efficient Indexing in Spatiotemporal Databases*, Proceedings 10ª Conferenza Internazionale SSDBM, Capri, Italia, 1998 pp. 123-132

[23] N. Tryfona, T. Hadzilacos, *Logical Data Modeling of Spatio-Temporal Applications: Definition and a Model*, Chorochronos Project – Technical Report Series, CH-97-03, 1997

[24] N. Tryfona, C.J. Jensen, *Conceptual Data Modeling for Spatiotemporal Applications*, Chorochronos Project -Technical Report Series, CH-98-08, 1998
[25] J. Van den Bussche D. Van Gucht, G. Vossen, *Reflecting Programming in the Relational Algebra*, Journal of Computer and System Sciences, 1996, Vol. 52, N. 3, pp.537-549
[26] P. Violi, *Significato ed esperienza*, Bompiani, Milano, 1997
[27] A. G. Wilson, *Geography and the environment: system analytical methods* – Wiley, 1981
[28] G. Wiederhold, *Mediators in the Architecture of Future*, Information Systems, IEEE Computer 25(3), 1992, pp. 38-49
[29] M.F. Worboys, *A Unified Model for Spatial and Temporal Information*, in *The Computer Journal*, Vol. 37, n° 1, 1994, pp. 27-34

Quinta
Conferenza
Nazionale
di Statistica

WORKSHOP

Esperienze di riorganizzazione di processi statistici nel Sistema statistico nazionale

coordinatore

Luigi Marini
responsabile dell'ufficio di statistica
del Ministero della giustizia

Roma 15, 16, 17 novembre 2000
Palazzo dei Congressi
Piazzale Kennedy - EUR

ISTAT

WORKSHOP

Esperienze di riorganizzazione di processi statistici nel Sistema statistico nazionale

coordinatore

Luigi Marini
responsabile dell'ufficio di statistica
del Ministero della giustizia

WORKSHOP

Esperienze di riorganizzazione di processi statistici nel Sistema statistico nazionale

La documentazione dei processi e dei prodotti statistici: strumenti, problematiche e prospettive

R. Carbini, V. De Santis, M. Fortini, M. Signore

QUINTA CONFERENZA NAZIONALE DI STATISTICA

ESPERIENZE DI RIORGANIZZAZIONE DI PROCESSI STATISTICI NEL SISTEMA STATISTICO NAZIONALE

LA DOCUMENTAZIONE DEI PROCESSI E DEI PRODOTTI STATISTICI: STRUMENTI, PROBLEMATICHE E PROSPETTIVE

R. Carbini, V. De Santis, M. Fortini, M. Signore

Istat

versione provvisoria

Roma, 16 novembre 2000

1. Introduzione

Nel presente lavoro viene descritta la strategia definita dall'Istat per il miglioramento della qualità dei processi e dei prodotti statistici del Sistema statistico nazionale.

La prima fase di tale strategia, che si è appena conclusa, è stata orientata a standardizzare sia l'attività di produzione che l'attività di documentazione. A tal fine sono stati messi a punto due strumenti di supporto: il primo è una lista di verifica, disponibile sia su supporto cartaceo che informatizzato, la quale consente di documentare, seguendo uno schema strutturato e standard, le caratteristiche di processo e la qualità dei dati prodotti; il secondo è un manuale in linea che si caratterizza come una raccolta di migliori pratiche correnti per la progettazione, l'esecuzione ed il controllo di una rilevazione statistica. .

Il lavoro è articolato come segue. Nel paragrafo 2 vengono svolte alcune considerazioni sulle diverse finalità della documentazione e sui diversi approcci che possono essere seguiti nell'avviare un'attività sistematica di documentazione. Le caratteristiche principali della lista di verifica vengono descritte nel paragrafo 3, in particolare si descrive come la lista possa essere efficacemente utilizzata sia per descrivere diverse tipologie di processi produttivi, sia per produrre note metodologiche differenziate in funzione delle esigenze conoscitive di diverse tipologie di utenti. Per facilitare, in termini di efficienza ed efficacia, l'attività di documentazione e per consentire di riutilizzare note metodologiche già prodotte, è stato realizzato uno strumento operativo (la versione informatizzata della lista di verifica), come descritto nel paragrafo 4. Le caratteristiche del relativo software sono invece riportate nel paragrafo 5. Nel paragrafo 6, viene sinteticamente illustrato il manuale "linee guida per rilevazioni statistiche". Le considerazioni conclusive vengono svolte nel paragrafo 7.

Il manuale in linea e la lista di verifica sono consultabili sul sito WEB dell'ISTAT all'indirizzo http://www.istat.it/metadati/lineeg/lg.html

2. Documentazione: finalità e possibili approcci

L'attività di documentazione può essere finalizzata al conseguimento di due obiettivi principali:

- Garantire e migliorare la qualità
- Migliorare la comprensione e l'uso dei dati statistici

Con riferimento al primo obiettivo, citiamo come la disponibilità di documentazione, in particolare quella sui processi, è un pre-requisito per un'attività di monitoraggio finalizzata ad assicurare il conseguimento ed il mantenimento di standard qualitativi. In tal senso la documentazione sui processi può essere utilizzata in primo luogo come base per controllare il processo durante la sua esecuzione, e poi come uno strumento per formare nuovo personale, per garantire che non si disperda la conoscenza acquisita nei casi di mobilità del

personale, ecc. In modo analogo la documentazione sui processi esistenti può essere utilizzata per progettare nuove indagini con caratteristiche simili. Poiché la documentazione mette in evidenza i punti di forza e di debolezza degli attuali processi produttivi, costituisce anche un punto di partenza per apportare modifiche ai processi finalizzate al miglioramento della qualità dei dati prodotti.

Un altro importante obiettivo della documentazione consiste nel mettere gli utenti (sia interni che esterni al Sistema statistico nazionale) in condizione di utilizzare nel modo migliore i dati disponibili. Questo implica in primo luogo la disponibilità di documentazione che consenta agli utenti di conoscere le fonti informative disponibili sui fenomeni di interesse e di discriminare tra le diverse fonti, (metadati sul contenuto informativo: fenomeni, unità di analisi, definizioni , classificazioni,…). In secondo luogo gli utenti dei dati statistici, in particolare quelli che vogliono effettuare ulteriori analisi o ricerche, necessitano di informazioni che consentano loro di valutare la qualità dei dati che intendono utilizzare (in particolare misure relative all'accuratezza dei dati come le stime degli errori campionari e di quelli non campionari, e in aggiunta indicatori di qualità relativi al processo, come ad esempio i tassi di risposta).

Occorre tuttavia osservare che la documentazione del processo e la misurazione della qualità sono operazioni che richiedono tempo e che possono anche essere vissute dai responsabili dei processi come azioni che ritardano la loro attività principale costituita dalla produzione di informazione. Per questi motivi devono essere predisposti strumenti che facilitino il più possibile l'attività di documentazione rendendola al tempo stesso più veloce. L'Istat ha avviato da alcuni anni una strategia per la documentazione basata sulla realizzazione di sistemi informativi statistici. In particolare è in fase di completamento lo sviluppo del Sistema informativo di documentazione delle indagini – SIDI, il quale gestisce metadati e indicatori di qualità standard sui processi di produzione, ed è stato effettuato uno studio di fattibilità per la realizzazione di un sistema informativo per la documentazione delle unità di analisi, variabili e classificazioni osservate nelle indagini e nei sistemi informativi statistici- SDOSIS.

Per avviare un'attività di documentazione più standardizzata e più sistematica anche all'interno del Sistan, è stato deciso di predisporre in una prima fase degli strumenti di supporto che fossero più flessibili e quindi più facilmente adattabili alle diverse realtà esistenti all'interno del Sistema. In questo modo si è privilegiato l'avvio e il radicamento di un'attività di documentazione, favorendo lo scambio di esperienze, l'uso di metodologie comuni e la diffusione di un linguaggio comune all'interno del Sistan. Per rispondere a queste esigenze è stata predisposta la lista di verifica per la documentazione dell'attività statistica, oltre al manuale in linea. L'uso di una lista comune da parte di enti diversi aumenta, infatti, la confrontabilità delle note metodologiche prodotte, facilitando così l'utenza anche al di fuori del Sistema statistico nazionale e garantendo il requisito della trasparenza.

Dopo una fase di diffusione di tali strumenti e di consolidamento dell'attività di documentazione, è previsto l'avvio di uno studio di fattibilità per valutare le possibilità e le modalità di estensione dei sistemi informativi di documentazione sopra citati, in particolare di SIDI, all'intero Sistema.

3. La lista di verifica: caratteristiche principali e possibili utilizzazioni

Obiettivi

La "Lista di verifica per la documentazione dell'attività statistica", disponibile sia su supporto cartaceo che informatizzato, consiste in un insieme di domande a risposte aperte per ciascuna fase di un processo statistico e si caratterizza come uno strumento che consente di soddisfare diversi obiettivi.

In primo luogo è uno strumento di supporto per i responsabili di un processo di produzione di informazione statistica per documentare, seguendo uno schema strutturato e standard, le modalità di progettazione ed esecuzione del processo e gli indicatori di qualità calcolati. A tal fine la lista è stata strutturata in modo tale da consentire la documentazione di diverse tipologie di processi produttivi. Inoltre poiché i destinatari della documentazione sono una pluralità di soggetti con esigenze conoscitive e conoscenze statistiche anche molto differenziate, sono stati individuati dei sottoinsiemi di informazioni, attraverso una pre-selezione dei quesiti della lista, che consentono di produrre note metodologiche più o meno dettagliate e quindi più rispondenti alle esigenze conoscitive dei diversi profili di utenza.

In secondo luogo, la lista può essere utilizzata come una guida per l'attività di controllo della qualità sia in fase di progettazione sia durante l'esecuzione del processo. In fase di progettazione, infatti, il responsabile di un processo può prendere visione in modo schematico dei possibili controlli che possono essere messi in atto per ciascuna fase del processo produttivo ed effettuare le scelte ritenute più idonee. In fase di validazione dei risultati, invece, il responsabile è guidato nel controllare la qualità delle diverse fasi e nel calcolare misure relative alle possibili fonti di errore. Chiaramente per approfondimenti sulle tecniche di controllo della qualità e sugli stimatori più appropriati si deve ricorrere a testi specifici, oppure in prima istanza al manuale in linea.

La lista di verifica è stata sottoposta a verifica e applicata a differenti contesti produttivi nell'ambito delle attività del Gruppo di lavoro per la sperimentazione delle linee guida e degli standard per la produzione statistica degli uffici di statistica degli enti del Sistan[1], istituito su mandato del COMSTAT[2]. La sperimentazione, effettuata con soggetti rappresentanti realtà diverse del Sistema statistico nazionale, è stata finalizzata alla messa a punto della lista e all'adeguamento alle esigenze specifiche degli enti del Sistan, come descritto alla fine del capitolo. Tale adeguamento ha riguardato, per le rilevazioni statistiche, anche il manuale in linea.

[1] Il Gruppo, istituito con delibera presidenziale n.44/P del 18 gennaio 2000, ha chiuso i suoi lavori a giugno 2000

[2] Comitato di indirizzo e coordinamento dell'informazione statistica

Contenuto

La lista di verifica consente di documentare tutti gli aspetti rilevanti dei diversi processi produttivi di informazione statistica. In particolare sono comprese tutte le tipologie di processo presenti in ambito Sistan, ovvero: rilevazioni dirette (su persone, imprese, istituzioni), rilevazioni da fonte amministrativa (quando i dati amministrativi non siano patrimonio dell'ente titolare del processo); rilevazioni miste (quando le due precedenti modalità di acquisizione dei dati coesistono); elaborazioni di dati statistici (nel caso in cui vengano riutilizzati dati statistici provenienti da altre rilevazioni e/o elaborazioni); elaborazioni di dati amministrativi (nel caso in cui i dati siano in possesso dell'ente in ragione della propria attività istituzionale).

La lista è suddivisa in blocchi di quesiti (a risposte aperte) che identificano gli aspetti principali in cui può essere scomposto un processo produttivo. Chiaramente alcuni blocchi di quesiti sono di carattere generale e quindi applicabili indipendentemente dalla tipologia di processo, mentre altri sono specifici per ciascuna tipologia.

Per semplicità di esposizione si descrivono dapprima in modo sintetico i blocchi di quesiti relativi alle rilevazioni dirette, alcuni dei quali sono comuni anche alle altre tipologie, e successivamente si evidenziano alcuni aspetti particolari che riguardano le rilevazioni di fonte amministrativa e le elaborazioni.

Per le rilevazioni dirette sono previsti blocchi di quesiti sui seguenti aspetti:

- Dati identificativi del processo
- Progettazione del processo che include gruppi di domande specifiche su: obiettivi conoscitivi; popolazione oggetto di studio e unità di analisi; variabili e classificazioni; questionario; verifica preventiva del processo (sperimentazioni e indagine pilota).
- Esecuzione del processo suddivisa in: caratteristiche generali del processo (periodicità ed esaustività); caratteristiche del disegno campionario; unità di rilevazione, archivi e liste di riferimento; acquisizione dei dati; revisione manuale; codifica; registrazione; revisione automatica o semiautomatica; elaborazione dei dati; metodi di stima.
- Validazione e indicatori di qualità che comprende: rilevanza e tempestività; errori campionari; errori di copertura ed errori di lista; mancate risposte totali; mancate risposte parziali; errori di risposta ed errori di misurazione; validazione delle stime prodotte mediante confronti interni e con altre fonti.
- Vi sono quindi tre blocchi di quesiti relativi rispettivamente a: diffusione dei risultati; ristrutturazioni o riprogettazioni per migliorare la qualità; riferimenti alla documentazione disponibile.

Con riferimento alle rilevazioni di fonte amministrativa, è stato sviluppato un blocco di quesiti relativi alla fase di acquisizione dei dati, che consente di descrivere sia il processo amministrativo alla base della raccolta dei dati (archivi amministrativi utilizzati, informazioni relative agli atti amministrativi all'origine dell'archivio, controlli di qualità effettuati dall'ente titolare del processo amministrativo) sia le analisi svolte per l'uso a fini statistici dei dati amministrativi (valenza statistica dell'archivio amministrativo, modalità di trasferimento per fini statistici delle informazioni raccolte con l'atto amministrativo).

Per quanto riguarda le elaborazioni, sono stati predisposti quesiti differenziati distinguendo le elaborazioni su dati statistici dalle elaborazioni su dati amministrativi in possesso dell'ente. Nel secondo caso, infatti, viene data indicazione di rispondere a tutti i quesiti predisposti per le rilevazioni di fonte amministrativa con l'avvertenza che, in questo caso specifico, l'ente titolare del processo amministrativo e quello dell'elaborazione coincidono. Nel caso, invece, di elaborazioni su dati statistici sono stati sviluppati una serie di quesiti ad hoc che consentono di descrivere le fonti utilizzate, le metodologie di elaborazione dei dati di base e di integrazione tra le fonti, gli studi e gli eventuali interventi correttivi per migliorare la qualità dei dati di base.

Profili di utenza

Come è stato già accennato, gli utenti dell'informazione statistica hanno esigenze conoscitive diversificate e hanno bisogno di informazioni che consentano loro di utilizzare correttamente i dati. Ciò richiede la produzione di note metodologiche appropriate a ciascuna tipologie di utenti, le quali contengano una opportuna selezione degli aspetti metodologici ritenuti salienti per quella tipologia e che devono essere associate alla diffusione dei dati e delle ricerche, sia internamente che al di fuori del Sistan.

Per articolare la lista di verifica in funzione delle diverse esigenze informative sono stati individuati dei sottoinsiemi di quesiti della lista in base a due modalità di classificazione: a) tipologie di utenti, con le quali si identificano diverse categorie di destinatari della documentazione; b) criteri di valutazione dei quesiti.

Nel primo caso, la popolazione degli utenti è stata suddivisa in tre categorie, a seconda del grado di specializzazione e di conoscenze statistiche e del possibile uso dei dati finali, individuando i seguenti profili informativi:

- utenti esterni finali generici
- utenti esterni finali specializzati
- utenti Sistan

Per i criteri di valutazione è stata scelta una scala a tre livelli:

- molto rilevante
- abbastanza rilevante
- poco rilevante

In corrispondenza di ciascuna tipologia di utenza, i quesiti della lista di verifica sono stati filtrati sulla base delle modalità del criterio di valutazione, consentendo in tal modo la selezione di sottoinsiemi informativi. Di seguito si riporta una descrizione sintetica dei risultati ottenuti, mentre per una illustrazione approfondita si rimanda alla relazione conclusiva del Gruppo di lavoro già citato e al relativo allegato "Contributi".

La modalità “molto rilevante” è stata assegnata ai quesiti considerati irrinunciabili per uno specifico profilo di utenza: in questo modo si individua l’insieme minimale di informazioni da fornire ad una determinata tipologia di utenza, senza incorrere in una omissione di informazioni. La modalità “abbastanza rilevante” consente di ampliare il contenuto informativo, per ciascun profilo. Attraverso i quesiti cui è stata attribuita tale modalità si identifica l’informazione aggiuntiva che può essere fornita nel caso in cui si voglia predisporre una nota metodologica più dettagliata rispetto a quella ottenuta sulla base dei soli quesiti “molto rilevanti”. Una considerazione diversa deve essere fatta per i quesiti classificati come “poco rilevanti” per il corrispondente profilo di utenza In questo caso, infatti, è preferibile non includere queste informazioni nelle note metodologiche destinate a quei profili, in quanto l’informazione aggiuntiva potrebbe pregiudicare la fruibilità del testo. Un tipico caso è quello dell’utenza generica per la quale si raccomanda di produrre note sintetiche (usando solo i quesiti segnalati come “molto rilevanti”) o note dettagliate (aggiungendo i quesiti segnalati come “abbastanza rilevanti”), ma si sconsiglia di aggiungere informazioni ritenute “poco rilevanti” perché sono riferite ad aspetti che risultano essere troppo tecnici o troppo operativi.

Chiaramente esiste una gerarchia tra i profili di utenza in quanto ciò che è rilevante per l’utente generico lo sarà a maggior ragione per l’utente specializzato e per l’utente appartenente al Sistan.

Per l’utente SISTAN si è stabilito di assegnare sempre il punteggio “molto rilevante”, in quanto tutti i quesiti sono stati ritenuti necessari per una documentazione accurata del processo e della qualità dei dati che consentisse uno scambio informativo all’interno del Sistema garantendo il requisito della riproducibilità del processo. In generale, l’utente Sistan potrà essere interessato a tutti gli aspetti sottesi alle attività di ricerca e sarà comunque in grado di selezionare da solo, se necessario, l’informazione di maggiore interesse per i propri scopi.

Una particolare attenzione è stata invece posta nel differenziare il contenuto informativo delle note metodologiche indirizzate alle altre due tipologie di utenti esterni.

L’utenza generica, che costituisce il raggruppamento più numeroso, è composta esclusivamente da persone esterne al Sistema e comprende tutti coloro che indistintamente fruiscono delle informazioni statistiche (operatori, studenti, giornalisti, ecc.) e che fondamentalmente necessitano di poche informazioni, ma significative e intelligibili, sulle principali caratteristiche delle produzioni statistiche e il loro livello di attendibilità. Il sottoinsieme dei quesiti irrinunciabili per l’utenza generica assume un particolare rilievo in quanto individua di fatto le informazioni metodologiche fondamentali e al tempo stesso di facile comprensione che non richiedono conoscenze statistiche approfondite. Un’ulteriore utilizzazione di questo insieme di quesiti potrebbe essere quella di costituire il set minimale di informazioni con cui accompagnare le ricerche anche per quegli enti che, effettuando diverse attività e disponendo però di poche risorse, si trovino nella difficoltà di approfondire la stesura delle note metodologiche e necessitino di una selezione degli aspetti metodologici essenziali.

L’utente specializzato invece comprende i ricercatori e gli studiosi in generale che possono essere interessati ad effettuare ulteriori studi rielaborando i dati prodotti da una o più fonti informative e che

possiedono competenze tecniche e metodologiche. Di conseguenza per tale profilo sono stati selezionati tutti quei quesiti che fornissero informazioni utili per un corretto uso dei dati e che facessero emergere eventuali limitazioni per il loro utilizzo, come ad esempio la presenza di fattori distorsivi o l'interruzione di una serie storica. In generale vengono descritti tutti gli aspetti metodologici inerenti le caratteristiche dei dati, il loro trattamento e i relativi indicatori di qualità. Vengono invece trascurati o messi in secondo piano tutti gli aspetti più inerenti al processo produttivo in senso stretto che assumono rilevanza solo per l'utente Sistan, interessato ad effettuare confronti tra processi o a riprodurre in tutto o in parte un processo produttivo.

Sperimentazione

Come già detto, la lista di verifica è stata oggetto di sperimentazione su di un paniere di processi produttivi condotti in ambito Sistan e rappresentatitivi di differenti tipologie. Le note metodologiche prodotte in tale occasione sono contenute nel documento interno "Contributi" allegato alla relazione conclusiva del Gruppo di lavoro[3].

In tale fase di sperimentazione della lista di verifica sono emerse diverse esigenze che hanno portato a modifiche e a integrazioni alla lista stessa. Le modifiche di cui è stata oggetto la lista di verifica hanno riguardato, oltre alla individuazione dei diversi profili di utenza di cui si è già parlato, estensioni del contenuto informativo della lista per renderla più aderente alle diverse realtà presenti nel Sistan, la classificazione del suo contenuto informativo per tipologia di processo produttivo, nonché la predisposizione di una versione informatizzata per renderne più agevole e più flessibile l'uso (cfr. paragrafo 4).

Le principali integrazioni al contenuto informativo iniziale della lista di verifica hanno riguardato in modo particolare l'introduzione di diversi quesiti per descrivere le rilevazioni di fonte amministrativa e di alcuni blocchi di quesiti relativi ai processi di elaborazione, sia nel caso del trattamento di dati amministrativi in possesso dell'ente, sia nel trattamento di dati statistici. Ciò ha costituito una significativa innovazione della struttura iniziale della lista di verifica.

Un'altra esigenza che è stata espressa da parte del gruppo di lavoro è stata quella di dotare la lista di verifica di un glossario contenente le definizioni dei principali termini utilizzati. Il glossario consente non solo di chiarire il significato di alcuni termini, ma anche di standardizzare il linguaggio in ambito Sistan. E' stato scelto di definire solo i termini principali, in quanto la lista di verifica contiene già al suo interno alcune note esplicative che guidano nella compilazione della lista, quali ad esempio la differenza tra indagine pilota e test preventivo di una procedura. Per una descrizione più dettagliata si rimanda invece al manuale in linea nel quale vengono trattati molti concetti richiamati nella lista. E' importante sottolineare che le definizioni di elaborazione e di rilevazione (o indagine), riportate nel glossario, sono quelle adottate ufficialmente dal Programma statistico nazionale (PSN), mentre le altre definizioni sono tratte dal manuale operativo del SIDI e quindi utilizzate per la documentazione dei processi all'interno dell'Istat. Il glossario contiene inoltre i riferimenti normativi che regolamentano l'obbligo di risposta e che definiscono i dati sensibili.

[3] Gli enti che hanno documentato uno o più processi produttivi sono i seguenti: Istituto Tagliacarne; Ministero dell'Industria, del commercio e dell'artigianato; Regione Toscana; Unioncamere.

4. Informatizzazione della lista di verifica: adeguatezza eusabilità

La versione informatizzata della lista di verifica consiste in uno strumento che risponde ai criteri di:

- adeguatezza
- usabilità

Nell'accezione qui adottata, l'adeguatezza è intesa nel senso di favorire il raggiungimento degli obiettivi, che nel contesto in esame sono la verifica e la documentazione del processo; l'usabilità è la caratteristica di facilitare considerevolmente, rispetto allo strumento cartaceo, l'utilizzo della stessa lista di verifica. Di fatto lo strumento informatizzato permette di sperimentare la lista di verifica su un insieme variegato e completo di realtà in ambito Sistan.

I criteri di adeguatezza e usabilità che hanno ispirato l'ideazione e quindi la progettazione e la realizzazione della versione informatizzata della lista di verifica sono, naturalmente, tra loro interconnessi.

4.1 L'adeguatezza

Sia che l'obiettivo riguardi la verifica del processo o la sua documentazione, lo strumento informatizzato della lista di verifica può ritenersi adeguato (rispetto agli obiettivi in questione) ove permetta di verificare e/o documentare in modo chiaro ed efficace (anche con riferimento al tempo impiegato) il processo.

La struttura che sottende la lista di verifica è il risultato di un incrocio di cinque tipologie di processo:

- Rilevazione diretta
- Rilevazione da dati amministrativi in possesso di altri enti
- Rilevazione mista
- Elaborazione di dati amministrativi in possesso dell'ente titolare del processo
- Elaborazione di dati statistici

e cinque profili informativi, ossia profili con diverso dettaglio dei quesiti

- Profilo A, struttura sintetica destinata ad un utente finale generico
- Profilo B, struttura dettagliata destinata ad un utente finale generico
- Profilo C, struttura sintetica destinata ad un utente finale specializzato
- Profilo D, struttura dettagliata destinata ad un utente finale specializzato
- Profilo E, struttura completa destinata ad un utente Sistan

cui occorre aggiungere il profilo per la predisposizione personalizzata del documento. Ovviamente, ciascuno dei profili risponde sia alle esigenze di verifica del processo, sia a quelle di documentazione,

quest'ultima pensata anche in un'ottica di trasparenza dei dati prodotti, ovvero come strumento che ne accompagna l'utilizzo da parte dell'utente esterno.

Come descritto in precedenza, il numero di quesiti (v. Tav. 1) della lista di cui consta ciascun profilo informativo è variabile. Più in dettaglio si passa da un minimo di 16 quesiti ad un massimo di 117. Di conseguenza, pensare alla documentazione di un processo di produzione statistica disponendo solo dello strumento cartaceo implica non solo l'esatta individuazione della tipologia di processo ma anche, da parte di colui che effettua la documentazione, la lettura di un intero documento di cui solo parte risulta coerente con quel processo. In altre parole, lo strumento cartaceo non consente di filtrare altrettanto efficacemente rispetto a quello informatizzato i soli quesiti pertinenti al livello di documentazione che si intende fornire in relazione allo specifico processo di cui si è titolari e/o del destinatario della documentazione (profilo di utenza).

La traduzione a livello informatico dei concetti sinteticamente esposti consente di definire adeguato lo strumento informatizzato che consenta di utilizzare efficacemente la lista di verifica in ambito Sistan, ovvero sulla base delle tipologie di processo contemplate e sui profili informativi individuati.

Tav. 1 – NUMERO DI QUESITI PER PROCESSO E PER PROFILO INFORMATIVO

	PPS1	PPS2	PPS3	PPS4	PPS5
LC	103	90	117	52	37
UG	23	20	24	17	16
UGD	40	30	42	26	22
US	53	45	58	33	27
USD	84	74	98	48	33

Legenda

PPS1 = Rilevazione diretta; PPS2 = Rilevazione da fonte amministrativa; PPS3 = Rilevazione mista; PPS4 = Elaborazione su dati raccolti dalla propria amministrazione; PPS5 = Elaborazione su dati statistici. LC = Lista completa; UG = utente finale generico: tutti i quesiti con punteggio 1; UGD = utente finale generico, documentazione dettagliata: tutti i quesiti con punteggio 1 e 2; US = utente finale specializzato: tutti i quesiti con punteggio 1; USD = utente finale specializzato, documentazione dettagliata: tutti i quesiti con punteggio 1 e 2

4.2 L'usabilità

Il successo delle operazioni che coinvolgono sistemi complessi, quale è la realtà del Sistan, è funzione del grado di usabilità dei prodotti predisposti. Infatti, l'usabilità si definisce come la capacità di un prodotto di agevolare uno specifico utente nel raggiungimento di specifici obiettivi con efficacia, efficienza e soddisfazione in uno specifico contesto di utilizzo. Il contesto di utilizzo della lista di verifica è costituito dalla produzione della documentazione dei processi effettuata dai soggetti facenti parte del Sistema statistico nazionale.

Nell'ambito Sistan lo specifico utente è rappresentato dal responsabile del processo di produzione statistica, che tra i suoi compiti ha anche quello di controllare il processo e al termine dello stesso, di documentare e validare i prodotti.

Come accennato, la lista di verifica è strutturata secondo la tipologia del processo produttivo e secondo il profilo informativo, e consiste in un insieme variabile di quesiti pertinenti. Operativamente, il titolare del processo di produzione statistica può:

- controllare il processo, in fase di esecuzione, attraverso la lista di verifica
- documentare il processo contestualmente all'esecuzione delle fasi di cui consta

e quindi,

- disporre di una documentazione di livello informativo minimo standardizzato (utilizzabile per esempio come nota metodologica in appendice alla pubblicazione dei propri prodotti destinati ad una utenza generica)
- produrre, a partire dalla documentazione finalizzata all'utenza generica, una documentazione di livello informativo superiore, ad esempio finalizzata alle esigenze informative di utenti specializzati

In questo scenario, la realizzazione del software della lista di verifica, con funzioni di:

- supporto alla compilazione
- memorizzazione
- modifica di output specifici
- produzione di output specifici

risulta fondamentale per il conseguimento degli obiettivi prefissati con la documentazione standardizzata.

Una caratteristica che incrementa notevolmente il grado di efficienza del prodotto è costituita dall'adozione di una struttura ad albero di decisione per l'identificazione del tipo di processo, fatto in sé determinante ai fini della individuazione dei quesiti pertinenti.

L'albero di decisione è una tecnica di rappresentazione del processo decisionale che ha come obiettivo la determinazione del percorso migliore per il conseguimento dell'obiettivo. Inoltre, costituisce anche una tecnica di classificazione automatica che, basandosi su dati raccolti in precedenza, consente la corretta classificazione delle entità in esame mediante una sequenza di passaggi logici. La tecnica è in grado di risolvere problemi anche complessi di classificazione, a condizione di disporre di sufficienti dati per la sua costruzione. Il risultato è la scomposizione del processo in nodi di classificazione, costituiti da quesiti o espressioni che possono essere facilmente compresi dagli utenti.

Inoltre, solo una versione informatizzata può, rispetto a strumenti di tipo cartaceo, offrire all'utente percorsi di compilazione ottimali, per mezzo di quesiti le cui risposte guidano verso un determinato percorso di compilazione.

Infine, grazie all'integrazione con MS Excel è possibile produrre una notevole varietà di output, dalla stampa, al formato testo per la produzione di documenti Word, al formato .xls per l'uso del formato tabellare, ecc., e ciò permette un ulteriore livello di flessibilità e di riutilizzo della documentazione prodotta. Infatti in

tutti di casi di processi di produzione non occasionali, che costituiscono la grande maggioranza della produzione statistica del Sistan, la documentazione di un processo può essere riutilizzata aggiornando semplicemente la parte relativa agli indicatori (e naturalmente le date di riferimento). Oltre a ciò i casi di riutilizzo della documentazione standardizzata, nella variegata realtà dei soggetti Sistan, possono essere molteplici. Il riutilizzo, oltre ai fini suddetti nei precedenti paragrafi, può essere ipotizzato come base, ad esempio, per la stesura di documenti programmatici, consuntivi, di piani operativi, ecc. Questo, ovvero il riutilizzo, è un elemento chiave in termini di usabilità, in quanto il prodotto è destinato all'utilizzo da parte dei responsabili di processo, ovvero di figure professionali il cui tempo-lavoro costituisce una risorsa preziosa.

5. Le tecnologie, la documentazione e il controllo dei processi

Le documentazioni di processo (e il loro utilizzo per le operazioni di controllo) si sono basate in ogni epoca sulle tecnologie disponibili per la loro predisposizione, memorizzazione ed utilizzo. In epoche preistoriche le documentazioni di processo (il come si fa), si basavano sulla memoria umana e sulla trasmissione orale. Successivamente, in epoca storica e fino ai nostri giorni, l'invenzione della scrittura e l'uso dei supporti tradizionali, tavolette, papiri, carta, ecc. ha permesso di estendere nello spazio e nel tempo le possibilità della trasmissione orale e delle capacità mnemoniche dell'uomo.

Oggi, nell'anno Duemila, in piena rivoluzione informatica, la tecnologia offre nuove opportunità per la predisposizione, la memorizzazione e l'utilizzo delle documentazioni di processo, e restringendo il campo al nostro settore di interesse, della documentazione dei processi di produzione statistica dei soggetti Sistan. Infatti, come già detto nel precedente paragrafo, i vantaggi della predisposizione di una versione informatizzata sono notevoli nei termini di adeguatezza e di usabilità rispetto alla tradizionale versione cartacea. Sono state, quindi, utilizzate le opportunità offerte dalla tecnologia odierna, enunciabili nei termini di:

- utilizzo di interfacce grafiche
- programmazione ad oggetti
- utilizzo degli standard "de facto" del mondo Sistan
- facile distribuzione del prodotto e delle documentazioni mediante Web
- implementazione di maschere di input flessibili
- produzione di output in più formati standard
- installazione semplificata

Sulla base dei termini suddetti, è stata realizzata la versione 1.0 del software "Verifica e documentazione standard" (VDS) (v. Fig.1). Il software utilizza interfacce grafiche ed è progettato per la gestione delle

operazioni di imputazione, di modifica, di memorizzazione e di produzione degli output fino a 150 processi, limite perfettamente rispondente alla produzione statistica dei vari soggetti Sistan.

Per l'implementazione è stato utilizzato il linguaggio VBA (Visual Basic for Applications). Il Visual Basic for Applications (VBA) è un ambiente di sviluppo ad oggetti integrato in diverse applicazioni, tra cui Microsoft Office 97 e 2000. Il VBA offre un insieme di strumenti di programmazione che sono stati utilizzati per sfruttare le potenzialità dell'applicazione ospite. E questo ha consentito un notevole risparmio nei tempi di sviluppo.

Fig.1 - Verifica e documentazione standard

Infatti lo sviluppo contestuale alla versione cartacea della lista di verifica di una versione informatizzata, riduce di molto i tempi tecnici di implementazione dell'intero progetto (dalla formalizzazione del Gruppo di lavoro alla consegna dei prodotti operativi sono intercorsi solo 10 mesi).

I software che includono il VBA, come Microsoft Excel, sono chiamati applicazioni "customer oriented", ossia applicazioni che possono essere modellate sulle necessità. Questo permette di sviluppare rapidamente le soluzioni le quali consentono, tra l'altro, un apprendimento facilitato da parte dell'utilizzatore finale. Ciò significa che il software VDS, oltre ad essere stato sviluppato rapidamente potrà anche essere rapidamente utilizzato e con una manutenzione minima.

La scelta del linguaggio VBA è motivata sia dal già detto elevato valore del rapporto costi/prestazioni, ma anche dall'alto grado di portabilità in ambito Sistan delle applicazioni sviluppate con il VBA. Infatti l'applicazione per il suo funzionamento richiede l'ambiente Windows 95/98/NT/2000 e la presenza di MS Excel 97 o successivi, applicazioni che costituiscono oggi lo standard "de facto" degli uffici di statistica titolari di produzione statistica. Grazie all'integrazione con l'applicazione ospite (utilizzo delle finestra di stampa e di salvataggio di MS Excel) è possibile la produzione di output ad alto contenuto grafico (stampa con le opzioni offerte dall'anteprima, formato testo per la produzione di documenti Word, formato Excel per

l'uso del formato tabellare, …). Inoltre il VBA costituisce una ottima base per l'eventuale sviluppo su Web grazie alla sua compatibilità con il VBScript. Una caratteristica del software realizzato è la sua estrema flessibilità. Infatti è predisposto per quadri di input/output variabili ed è stata implementata una importante funzione di personalizzazione dei quadri di input/output, per cui è possibile adattare la lista di verifica anche a particolari esigenze.

Tra l'altro, l'installazione risulta essere molto semplificata, in quanto il programma è contenuto in una cartella di lavoro di Microsoft Excel (vds.xls), e può essere semplicemente copiato su una directory e da lì avviato mediante un doppio click sulla sua icona.

Infine, il software consente la produzione di output predisposti per una facile trasmissione via e-mail dei vari tipi di documentazione, mediante l'utilizzo delle funzioni integrate in MS Office. Inoltre le sue ridotte dimensioni (768 Kb in versione espansa, 400 Kb in versione .zip) lo rendono ideale per la sua distribuzione su semplice floppy-disk (versione espansa) o per il download da Web (versione .zip).

6. Il manuale "linee guida metodologiche per rilevazioni statistiche"

L'obiettivo di questo strumento è quello di trasmettere le principali nozioni di base connesse alla progettazione e all'esecuzione dei processi identificati come rilevazioni, sia che la raccolta dei dati avvenga attraverso un questionario (raccolta diretta), sia derivi da fonti organizzate (basata su fonti amministrative).

I potenziali fruitori del manuale in linea sono costituiti sia da coloro che, senza essere statistici, si trovano ad essere coinvolti in produzioni di tipo statistico, sia dagli utenti finali dei dati prodotti dagli uffici di statistica in generale e dall'Istat in particolare. È infatti noto che una rilevazione necessita di un elevato grado di organizzazione e dell'apporto di numerose professionalità oltre quella dello statistico propriamente detto. Alle fasi di pianificazione e gestione concorrono numerose competenze diverse da quelle dell'esperto di statistica come, solo per citarne alcune, gli esperti del fenomeno oggetto di studio, i funzionari amministrativi, gli informatici e i responsabili degli enti territoriali. Ad un livello più operativo è anche utile menzionare ad esempio coloro che adempiono alle fasi di contatto dei rispondenti (rilevatori), al trasporto e alla revisione del materiale, alla registrazione su supporto informatico e così via.

Per questo motivo è di vitale importanza per la riuscita della rilevazione che tutte le figure coinvolte abbiano un vocabolario comune, una visione generale del processo al quale stanno contribuendo e la consapevolezza di quanto il risultato del lavoro di loro competenza incida sui diversi snodi del processo produttivo e sulla qualità dell'informazione prodotta. Ecco quindi perché potenziali fruitori del manuale sono innanzitutto coloro i quali, prevalentemente all'interno del Sistan, si trovino coinvolti a qualche titolo nelle fasi di disegno o produzione di una rilevazione, senza per questo possedere specifiche conoscenze di tipo statistico.

Il manuale può però risultare utile anche ad utenti che non siano direttamente coinvolti in produzioni statistiche, ma che siano tuttavia interessati all'argomento. Fra questi è importante annoverare l'utente finale

dell'informazione in quanto, se si accetta l'impostazione secondo la quale la qualità del prodotto (l'informazione nel nostro caso) è guidata dall'utente, diventa di sicuro centrale che questi sia dotato di strumenti di tipo formativo in modo che gli sia facilitata la lettura critica dei dati ad esso forniti dall'Istat o da qualsiasi altro ente del Sistan. Questo è tanto più vero quanto più si faccia riferimento ad utenti non professionali, come gli operatori nel settore delle imprese o i semplici cittadini, ai quali sempre di più si cerca di facilitare l'accesso all'informazione statistica non mediato dai mezzi di comunicazione di massa. Per queste considerazioni il manuale on-line, pur tenendo conto della necessità di sostenere un ulteriore sforzo di semplificazione e sintesi dei concetti esposti, può senza dubbio andare incontro a quanti desiderino abbinare all'informazione statistica ad essi fornita anche delle nozioni sul complesso mondo della pianificazione e dell'esecuzione delle rilevazioni.

Vale infine la pena di citare tra i potenziali fruitori gli studenti di statistica o quegli statistici i quali non siano mai stati direttamente coinvolti nei processi di reperimento, raccolta e validazione dell'informazione statistica nell'ambito della statistica ufficiale. Ad essi infatti il manuale propone, insieme a sezioni introduttive di agile consultazione, gli spunti bibliografici necessari per gli approfondimenti desiderati.

Il manuale, predisposto in formato html in modo da consentirne la consultazione direttamente via Internet, è organizzato in circa trenta diverse sezioni nelle quali si illustrano sia gli aspetti riguardanti la pianificazione di una rilevazione sia i temi concernenti le fasi operative. Come già detto, il taglio dato all'illustrazione dei vari argomenti è volutamente di base dato che lo strumento si rivolge ad un lettore non esperto con l'obiettivo di fornire un'introduzione ai temi trattati. Tuttavia la trattazione, che tiene conto sia dell'esperienza dell'Istat sia dei contributi di autorevoli esperti delle materie trattate, prevede, per le parti più operative, apposite sottosezioni rivolte ai responsabili di processo dove si forniscono raccomandazioni applicative finalizzate al conseguimento di risultati di qualità.

La forma ipertestuale utilizzata nella redazione è risponde all'obiettivo di rappresentare le relazioni esistenti fra le diverse fasi di una rilevazione piuttosto che l'approfondimento di temi specifici. La modularità del formato adottato permette comunque sia di inserire in futuro ulteriori sezioni finalizzate all'approfondimento, sia di aggiungerne altre sui nuovi argomenti che si riterrà utile descrivere. Un indice generale consente di accedere direttamente ad ogni sezione mentre i legami ipertestuali predisposti all'interno di ogni parte consentono al lettore di raggiungere direttamente gli altri argomenti ad essa attinenti. Sulla sinistra di ogni pagina è inoltre sempre disponibile una lista di collegamenti che puntano direttamente ai principali temi affrontati dal manuale, facilitando così le ricerche. Una pagina specifica è dedicata ai riferimenti bibliografici che, oltre ad essere accessibili dall'indice generale, sono facilmente raggiungibili da ogni citazione riportata nel testo.

7. Conclusioni

L'esperienza acquisita nel lavoro di definizione e sperimentazione della lista di verifica e del manuale in

linea suggerisce che tali strumenti possano essere utilizzati in modo efficace dal Sistema statistico nazionale per avviare un utile percorso di omogeneizzazione della documentazione e delle metodologie utilizzate. Al fine di promuovere l'uso di tali strumenti occorre tuttavia attuare una strategia che li ponga al centro di iniziative in grado di coinvolgere gli enti del SISTAN e creare condizioni favorevoli al radicamento degli standard documentali. Un elemento utile a questo proposito è certamente costituito dai requisiti di flessibilità e facilità di applicazione degli strumenti concepiti. La predisposizione della lista di verifica su supporto informatizzato viene incontro a tali istanze soprattutto per quanto riguarda la modularità rispetto alle tipologie dei processi e alle esigenze informative dei titolari e degli utenti la documentazione. Per quanto riguarda invece il manuale in linea, la sua disponibilità nel sito WEB dell'Istat e la sua caratteristica di documento in formato ipertestuale dovrebbero facilitarne l'accesso da parte degli utenti appartenenti al SISTAN. Inoltre è previsto che le note metodologiche che saranno via via predisposte dagli enti del SISTAN siano raccolte in un unico database consultabile via Internet in modo che la facilità di accesso all'informazione prodotta costituisca un incentivo all'adeguamento agli standard da parte degli enti che sarebbero portati a considerarsi anche come fruitori delle informazioni piuttosto che soltanto come fornitori.

Un successivo momento di estensione degli standard documentali prevede che le informazioni raccolte sui processi di produzione statistica permettano di effettuare confronti temporali fra più occasioni di indagine riferite al medesimo processo o confronti fra processi produttivi differenti. Queste esigenze non potrebbero essere soddisfatte con gli strumenti attualmente proposti, ma richiedono piuttosto la predisposizione di appositi sistemi informativi. A questo proposito è utile prevedere in prospettiva l'estensione al SISTAN delle esperienze sviluppate in questo ambito dall'Istat con il Sistema di Documentazione delle Indagini (SIDI), progettato in modo da consentire di documentare in modo standard i processi (ad esempio mediante un sistema di tesauri) e di gestire nel tempo sia la documentazione di processo che un set di indicatori di qualità dei dati.

Una ulteriore considerazione può essere fatta riguardo agli archivi finalizzati a scopi gestionali. Per questi, così come avviene per le rilevazioni e le elaborazioni, possono essere definiti gli standard di documentazione e i requisiti di qualità affinché possano essere utilmente sfruttati con finalità statistiche. Questo aspetto infatti, se considerato alla luce dell'importanza in termini quantitativi di tali fonti in ambito SISTAN, riveste un grande potenziale per la conoscenza di numerosi aspetti della realtà del nostro Paese e può costituire una delle linee di sviluppo future del lavoro finora svolto.

In un'ottica di sviluppo futuro deve essere anche collocata l'esperienza degli enti del SISTAN, dove le indagini statistiche vengono svolte con il contributo di più soggetti in esecuzione di rapporti contrattuali. In questo caso, infatti, gli enti hanno bisogno di proporsi con sempre maggiore incisività quali garanti della qualità dell'informazione prodotta affinché questa possa essere diffusa come informazione ufficiale. Proprio per questo il rapporto contrattuale deve essere impostato sulla base di capitolati che indichino con buon dettaglio le modalità di realizzazione delle specifiche fasi oggetto di contratto, nonché la relativa documentazione. In prospettiva quindi si può vedere uno sviluppo della lista di verifica proprio con l'obiettivo di facilitare la predisposizione di collaborazioni tra gli enti titolari dei dati prodotti e le parti che

prendono in carico specifici segmenti di produzione, definendo al meglio i requisiti di qualità da garantire e la documentazione da produrre.

Da ultimo occorre citare il ruolo che la lista di verifica e il manuale possono svolgere come strumenti di riferimento attorno ai quali predisporre l'attività di formazione ad enti del SISTAN riguardo alle tematiche del controllo di qualità dei dati e della documentazione dell'attività statistica. A regime tale attività formativa, assieme alla produzione di note metodologiche diffuse a corredo dei dati statistici (tramite Internet, in apposite pubblicazioni o in appendice alle pubblicazioni contenenti i dati), potrà divenire parte di una organica politica di diffusione della cultura statistica mirata a favorire sia la qualità e la trasparenza del processo di produzione che il corretto uso dei dati da parte dei loro utilizzatori.

Quinta
Conferenza
Nazionale
di Statistica

WORKSHOP

Sistemi distribuiti e architettura dei dati

coordinatore

Giorgio Ausiello
docente dell'Università di Roma "La Sapienza"

Roma 15, 16, 17 novembre 2000
Palazzo dei Congressi
Piazzale Kennedy - EUR

ISTAT

WORKSHOP

Sistemi distribuiti e architettura dei dati

coordinatore

Giorgio Ausiello
docente dell'Università di Roma "La Sapienza"

WORKSHOP

Sistemi distribuiti e architettura dei dati

Architettura e integrazione dei dati in sistemi distribuiti

M. Lenzerini

Architettura e integrazione dei dati in sistemi distribuiti

Maurizio Lenzerini
Dipartimento di Informatica e Sistemistica
Università di Roma ``La Sapienza''
lenzerini@dis.uniroma1.it

1. Introduzione

L'integrazione dei dati è uno degli aspetti più significativi ed importanti nella progettazione e nella realizzazione di sistemi informativi distribuiti. Negli ultimi anni, le problematiche relative all'integrazione sono studiate in diversi ambiti:

1) basi di dati e data warehouse, in cui sia l'integrazione degli schemi, sia l'integrazione dei dati ricoprono un ruolo centrale;
2) sistemi informativi cooperativi, in cui l'integrazione costituisce uno dei problemi da risolvere per consentire la cooperazione, ad esempio tramite la condivisione di informazione;
3) sistemi informativi globali, in cui sorgenti di informazioni eterogenee vengono messe in linea (l'esempio più significativo in questo senso è il Web) e se ne richiede un accesso integrato.

Lo scopo di questo documento è di sottolineare sia i risultati raggiunti dalla ricerca sull'integrazione negli ultimi anni, sia i problemi fondamentali che sono ancora in fase di studio e approfondimento.

Per *integrazione dei dati* intendiamo il problema di fornire un'interfaccia uniforme a diverse sorgenti di dati, allo scopo di consentire ad un cliente del sistema di focalizzare l'attenzione su *qual è* la sua esigenza informativa, piuttosto che su *come agire* per soddisfarla. Di conseguenza, il sistema di integrazione solleva il cliente dalla necessità di scoprire quali sono le sorgenti rilevanti per le sue esigenze, di interagire con ognuna di esse separatamente, e di raccogliere, selezionare, e fondere i dati che provengono dalle sorgenti stesse.

La ricerca sull'integrazione dei dati è stata molto attiva negli ultimi anni [Hull97,Levy99]. Si è ormai raggiunto un consenso generalizzato sulla architettura di base di un sistema di integrazione, e sui parametri fondamentali che ne caratterizzano le proprietà. In tale architettura, due tipi di moduli giocano un ruolo fondamentale: *wrappers* e *mediatori*. Lo scopo di un wrapper è di incapsulare una sorgente di dati, di accedervi secondo meccanismi predefiniti, e di presentare i dati della sorgente secondo uno specificato formato. Il ruolo di un mediatore è invece di collezionare, ripulire, e combinare i dati che provengono dai wrapper o da altri mediatori, al fine di soddisfare una ben specificata esigenza informativa.

L'integrazione dei dati può essere *virtuale* o *materializzata*. Nel primo caso il sistema agisce come interfaccia tra il cliente e le sorgenti, mentre nel secondo caso il sistema mantiene una vista replicata dei dati delle sorgenti. L'approccio virtuale è tipico dei sistemi informativi globali, mentre quello materializzato è adottato nei data warehouse. Nei sistemi che seguono l'approccio materializzato la gestione e l'aggiornamento delle viste è una problematica cruciale, che richiede tecniche specializzate. Recentemente, sono stati proposti sistemi prototipali che conciliano entrambi gli approcci [JLVV99,CDLNR98d].

Il seguito di questo documento è dedicato ad una sintetica analisi dei problemi cruciali che sono ancora allo studio nei progetti di ricerca concernenti l'integrazione dei dati.

2. Progetto e realizzazione di schemi e mediatori

Riferendoci all'architettura descritta in precedenza, la specifica e la realizzazione di mediatori è uno degli aspetti cruciali della progettazione di un sistema di integrazione, ed è affrontata con due filosofie alternative: *fondata sulle sorgenti*, e *fondata sulle query*. Nel primo caso, sia le query che le sorgenti vengono descritte come delle viste su uno schema di riferimento (detto schema globale, o schema mediato), e per ogni query, il corrispondente mediatore ha il compito di pianificare il processo di risposta alle query accedendo alle sorgenti. Nel secondo caso, sono i concetti dello schema mediato ad essere definiti come viste sulle sorgenti, e il mediatore associato ad una query espande la query stessa usando tali definizioni, così da ottenere il piano di accesso alle sorgenti. Molti sistemi prototipali attuali seguono la filosofia fondata sulle query, ma un confronto tra i due approcci rivela che la filosofia fondata sulle sorgenti offre numerosi vantaggi in termini di modularità del sistema e di riusabilità dei suoi componenti [Ullm97].

3. Pianificazione degli accessi alle sorgenti

Dalle osservazioni precedenti, risulta chiaro che l'aspetto fondamentale nel progetto di mediatori è la caratterizzazione del piano di accesso alle sorgenti che consente di ottenere la risposta ad una query. Questo problema è noto in letteratura come *rispondere alle query usando le viste*, ed è attualmente oggetto di indagine in numerosi progetti di ricerca [Levy99]. Lo studio si è concentrato su specifiche classi di query, come le query congiuntive (con o senza operatori di confronto), le query disgiuntive, le query con operatori di aggregazione (group-by), le query ricorsive, le query espresse in logica, e le query con espressioni regolari. Sono stati sviluppati diversi algoritmi e sono state fornite diverse caratterizzazioni della complessità computazionale del problema. Dato il crescente interesse per sorgenti semi strutturate (ad esempio, sorgenti espresse in XML), di particolare rilevanza è il caso di query con espressioni regolari. La maggiore difficoltà in questo ambito risiede nel fatto che con le espressioni regolari si può esprimere una forma limitata di ricorsione, e ciò pone interessanti quesiti sulla complessità computazionale e, addirittura, sulla decidibilità del calcolo della risposta alle query.

4. Sorgenti ad accesso limitato

In molte applicazioni che richiedono l'integrazione di dati, le sorgenti informative a disposizione pongono dei limiti sul tipo di query che possono trattare [Levy99]. Ad esempio, una sorgente accessibile solo con una interfaccia basata su form (comunemente usate nel Web) può rispondere solo a query in cui il valore di uno o più parametri deve essere esplicitamente fornito in ingresso. Questa caratteristica complica notevolmente il problema della risposta alle query usando le viste. Infatti, la pianificazione degli accessi alle sorgenti deve ora tenere conto dei vincoli di accesso, la cui presenza accentua il problema della esistenza di piani alternativi. Il sistema deve quindi assicurare un adeguato livello di qualità del piano scelto rispetto a quelli scartati.

5. Conversione e riconciliazione dei dati

L'integrazione di dati provenienti da sorgenti eterogenee richiede anche tecniche per la conversione e la riconciliazione dei dati. La complessità delle conversioni e delle riconciliazioni è dovuta a diversi problemi legati alla eterogeneità delle sorgenti, quali: differenze nella rappresentazione di uno

stesso oggetto, possibili errori nella codifica di proprietà di oggetti, possibili discrepanze nelle proprietà attribuite agli oggetti nelle varie sorgenti. Sebbene esistano diverse tecniche e strumenti ad hoc, utilizzati specialmente nelle realizzazioni di data warehouse, solo recentemente il problema della conversione e riconciliazione dei dati è stato affrontato con metodi rigorosi e scientifici [CDLNR00].

6. Sorgenti semi strutturate

Grande interesse sta suscitando il problema dell'integrazione nel caso in cui le sorgenti informative contengano dati semi strutturati. È probabile che il diffondersi di XML accentuerà questo interesse nei prossimi anni, e favorirà lo studio di tecniche che tengano conto del coesistere di sorgenti strutturate e semi strutturate. Tutti i problemi precedentemente descritti si riproporranno in una veste più complessa e articolata. Ad esempio, la scelta tra una architettura fondata sulle query ed una fondata sulle sorgenti dovrà tenere presente che nel nuovo contesto i meccanismi descrittivi delle sorgenti dovranno necessariamente essere più sofisticati. Analogamente, il processo di risposta alle query usando le viste dovrà adattarsi alle nuove esigenze poste dalla presenza di sorgenti semi strutturate. La presenza di vincoli di accesso alle sorgenti semi strutturate porrà ulteriori complicazioni, sia sulla modellazione di tali vincoli, sia sui meccanismi di accesso da utilizzarsi nella sintesi dei piani. Infine, la necessità di conversioni e riconciliazioni di dati strutturati e semi strutturati pone nuovi e stimolanti problemi, che necessitano di precise caratterizzazioni e approfondimenti.

Riferimenti

[CDLNR98d] Diego Calvanese, Giuseppe De Giacomo, Maurizio Lenzerini, Daniele Nardi, and Riccardo Rosati. Information integration: Conceptual modeling and reasoning support. In *Proceedings of the Sixth International Conference on Cooperative Information Systems (CoopIS'98)*, pages 280--291, 1998.

[CDLNR00] Diego Calvanese, Giuseppe De Giacomo, Maurizio Lenzerini, Daniele Nardi, and Riccardo Rosati. Data integration in data warehousing. International Journal of Cooperative Information Systems, 2000. To appear.

[Hull97] Richard Hull. Managing semantic heterogeneity in databases: A theoretical perspective. *Proceedings of the Sixteenth ACM SIGACT SIGMOD SIGART Symposium on Principles of Database Systems (PODS'97)*, 1997.

[JLVV99] Matthias Jarke, Maurizio Lenzerini, Yannis Vassiliou, and Panos Vassiliadis, editors. *Fundamentals of Data Warehouses*. Springer-Verlag, 1999.

[Levy99] Alon Y. Levy. Answering queries using views: A survey. Technical report, University of Washinghton, 1999.

[Ullm97] Jeffrey D. Ullman. Information integration using logical views. *Proceedings of the Sixth International Conference on Database Theory (ICDT'97)*, volume 1186 Lecture Notes in Computer Science, pages 19--40. Springer-Verlag, 1997.

WORKSHOP

Sistemi distribuiti e architettura dei dati

OLAP e data mining in ambienti distribuiti

D. Saccà e D. Talia

OLAP e Data Mining in Ambienti Distribuiti

Domenico Saccà e Domenico Talia

ISI-CNR
Via P. Bucci, Cubo 41C
87036 Rende (CS), Italy
Email : {sacca,talia}@si.deis.unical.it

1. Introduzione

L'informazione memorizzata nei calcolatori oggi è enorme ed tende a crescere sempre più rapidamente. Mentre fino ad alcuni anni fa il principale problema era la mancanza o la carenza di informazione, oggi e in futuro il principale problema è la sovrabbondanza di flussi di informazione da cui è molto difficile estrarre cosa è interessante e importante. Solo per citare alcuni esempi, la catena di negozi Wal-Mart negli Stati Uniti gestisce 20 milioni di transazioni al giorno utilizzando una base di dati di oltre 4 terabytes di memoria, il progetto sul Human Genoma ha collezionato alcune decine di terabyte di dati sul codice genetico umano, un database contenente un catalogo degli oggetti celesti provenienti da un osservatorio astronomico consiste di miliardi di immagini, il sistema di osservazione terrestre dei satelliti orbitanti costruito dalla NASA genera 50 gigabytes di immagini all'ora. La quantità di informazione su INTERNET cresce molto rapidamente: si calcola che ogni due ore un nuovo server WEB venga connesso alla rete, e si è passati da circa 5 milioni di documenti web nel 1995 a oltre 1 miliardo nel 2000 ma la quantità di informazione veramente utile rimane limitata: si usa dire *che il 99% dell'informazione è inutile al 99% degli utenti WEB*.

Questi volumi di dati rendono chiaramente poco utili i metodi tradizionali di analisi. Tali metodi, infatti, possono creare dei rapporti informativi sui dati analizzati, ma non sono in grado di analizzarne il contenuto e focalizzare l'attenzione sulle informazioni interessanti che essi contengono. L'esigenza per una nuova generazione di tecniche e strumenti capaci di assistere gli utenti in modo intelligente ed automatico nell'analizzare queste "montagne" di dati allo scopo di trovare conoscenza utile è pertanto effettiva. Queste tecniche e strumenti sono argomento di studio nel campo *Scoperta di Conoscenza nelle Basi di Dati (Knowledge Discovery in Databases)*. Un *Sistema di Scoperta della Conoscenza* è un sistema che trova conoscenza che precedentemente non aveva, cioè che non era implicita nei suoi algoritmi o esplicita nella sua rappresentazione della conoscenza di dominio. Il termine *Knowledge Discovery in Databases*, in breve *KDD*, è stato coniato nel 1989 per riferirsi all'intero processo di scoperta della conoscenza nei dati e per enfatizzare l'applicazione ad alto livello di particolari metodi di estrazione dei dati, noti sotto il nome di *OLAP* e *data mining* [3]. Le tecniche OLAP (On-Line Analytical Processing) permettono di organizzare i dati secondo relazioni multidimensionali che possono essere agevolmente manipolate per pervenire ai livelli di aggregazione desiderati. Le

tecniche di data mining consentono di individuare nei dati associazioni nascoste e correlazioni significative.

Negli ultimi anni la gran parte dei sistemi di OLAP e data mining sono stati progettati e realizzati su calcolatori sequenziali. Tuttavia, tali sistemi possono trarre ulteriori benefici dalle tecniche e dai sistemi emergenti nel settore del calcolo parallelo e distribuito, come i sistemi di *cluster computing* e di *grid computing* (griglie computazionali) [4], ed in generale dall'uso di reti ad elevate prestazioni. Il sistemi distribuiti di data mining hanno lo scopo di fornire algoritmi, strumenti ed applicazioni per l'analisi di dati distribuiti. Le sorgenti informative di grandi dimensioni sono spesso logicamente e fisicamente distribuiti e non possono essere trasferiti in un singolo sito. Inoltre, vi sono molte organizzazioni o imprese delocalizzate sul territorio che hanno bisogno di un approccio decentralizzato al supporto delle decisioni. Quindi, ogni volta che grandi moli di dati sono associate alla distribuzione geografica dei dati, dei sistemi e degli utenti è necessario integrare differenti tecnologie per implementare sistemi distribuiti per il data mining e la scoperta di conoscenza. Esempi significativi dove ambienti di data mining distribuito possono essere usati utilmente sono

- Grandi organizzazioni che hanno necessità di analizzare dati distribuiti;
- Mining di dati in possesso di differenti organizzazioni o laboratori;
- Mining di dati provenienti da sorgenti o sensori geograficamente distribuiti;
- Analisi di dati residenti in siti diversi con diverse tecniche di data mining per ottenere risultati in parallelo (meta-learning).

2. OLAP e Data Mining Distribuiti

Molto spesso il compito della implementazione di algoritmi su sistemi di calcolo distribuiti non è semplice ed è anche costoso. Quindi insieme alla scalabilità, altri importanti obiettivi sono l'indipendenza dall'architettura e la portabilità. Questo obiettivo di astrazione architetturale deve essere considerato nel progetto di applicazioni ed algoritmi di data mining insieme all'obiettivo dell'interoperabilità. Questi sistemi devono essere capaci di

1. gestire grandi data set di dati multidimensionali che possono essere geograficamente distribuiti;
2. analizzare differenti tipi di dati che possono essere non- o semi-strutturati;
3. offrire una facilità di uso tramite l'uso di modelli di *data mining programming* e di linguaggi di specifica di alto livello;
4. gestire gli aspetti di sicurezza e privatezza dei dati che sono connessi con la distribuzione delle risorse.

L'aspetto centrale in un sistema di OLAP e data mining distribuito è l'analisi distribuita dei dati che può essere realizzata secondo gli schemi seguenti:

- *Move Data*. I dati residenti su nodi di elaborazione remoti sono selezionati e trasmessi ad un nodo centrale (probabilmente un cluster) per l'elaborazione.
- *Move Models*. Ogni nodo elabora i dati localmente ed invia il modello predittivo ad altri nodi per ulteriori elaborazioni.

- *Move Results*: Ogni nodo elabora i dati fino ad ottenere un qualche risultato che invia ad altri nodi per completare l'analisi.

Alcuni sistemi di data mining distribuiti che sono stati recentemente proposti supportano combinazioni di questi schemi. Tuttavia gli aspetti architetturali e infrastrutturali sono approcciati da specifici punti di vista applicativi. Nei casi più semplici i dati dono detenuti da una stessa organizzazione e il loro formato e la loro disponibilità sono ben conosciuti. Nei casi più complessi, in cui si hanno dati eterogenei con più detentori alcuni requisiti richiesti dal data mining distribuito possono essere soddisfatti tramite l'integrazione di servizi di data mining su sistemi distribuiti.

Combinando i modelli ed i risultati che emergono da queste area tecnologiche è possibile supportare lo sviluppo di sistemi distribuiti di data mining e realizzare applicazioni innovative in questo settore. Ognuna delle due area coinvolte (data mining e distributed computing) può usare l'altra per ridurre i costi di sviluppo e di esecuzione e gli esperti di ognuna di esse possono concentrarsi sul proprio dominio di competenza sfruttando l'indipendenza dalle architetture sottostanti. Inoltre, le applicazioni distribuite di data mining potranno beneficiare degli sviluppi che si stanno ottenendo nella definizione di standard per il processo di DM come il modello CRISP-DM e il *predictive modeling markup language* (PMML) che è stato specificato per gestire differenti formati per i dati di input ed output.

Recentemente sono stati sviluppati alcuni sistemi di data mining distribuiti che operano su reti o cluster di workstation ed alcuni che operano su Internet. Inoltre è stata proposta una architettura distribuita per il data mining su griglie computazionali [1]. Alcuni di questi sistemi suono stati implementati in Java, come Papyrus [5] che usa Java aglets, JAM [9] che è basato sul paradigma degli agenti distribuiti implementati tramite Java applets e PaDDMAS [8] in cui ogni componente dell'architettura distribuita del sistema è realizzato come un oggetto in Java o CORBA e l'interfaccia con le sorgenti informative è realizzata tramite JDBC. Altri sistemi distribuiti di data mining sono Kensington Enterprise data mining suite [2], MASSON [7] e BODHI [6] che implementa un tipo di meta-learning detto *collettive data mining* basato sulla raccolta di modelli generati nei diversi nodi del sistema distribuito e sulla valutazione comparata della loro bontà predittiva.

3. Conclusioni

I sistemi di OLAP e data mining in ambiente distribuiti rappresentano una evoluzione delle tecnologie di knowledge discovery. Essi sono basati sul fatto che i dati possono essere inerentemente distribuiti in più siti che sono tra loro connessi da una rete di comunicazione e i dati stessi possono essere eterogenei. Questi sistemi offrono nuovi strumenti di scoperta della conoscenza che effettuano una analisi distribuita dei dati limitando i costi di comunicazione.

Naturalmente questi sistemi pongono nuovi problemi come la gestione della distribuzione dei dati e della computazione, i costi di comunicazione tra i nodi e il coordinamento e la gestione del processo decentralizzato di scoperta della

conoscenza. Tuttavia, anche a causa della diffusione di Internet e dei sistemi di elaborazione di rete, i sistemi distribuiti di data mining offrono nuove e significative opportunità per organizzazioni, imprese e ambienti scientifici che operano nel settore del knowledge management. Queste opportunità sono ancora maggiori se si considera il ruolo che gli ambienti distribuiti di data mining possono svolgere in settori strategici come l'analisi di dati disponibili su Internet come il Web mining, il Web personalization e il text mining.

Riferimenti bibliografici

1. Cannataro M. and Talia D., Parallel and Distributed Knowledge Discovery On The GRID: A Reference Architecture, *Proc. ICA3PP Conference*, Hong Kong, World Scientific Publ., 2000.
2. Chattratichat J., Darlington J., Guo Y., Hedvall S., Köler M. and Syed J., An architecture for distributed enterprise data mining. *HPCN Europe 1999*, *Lecture Notes in Computer Science,* **1593** (*Springer*, 1999) pp. 573-582.
3. Fayyad U.M. and Uthurusamy R. (eds.), Data mining and knowledge discovery in databases. *Communications of the ACM* **39** (1997).
4. Foster I. and Kesselman C., Globus : a metacomputing infrastructure toolkit. *International Journal of Supercomputing Applications* **11** (1997) pp. 115-128.
5. Grossman R., Bailey S., Kasif S., Mon D., Ramu A. and Malhi B., The preliminary design of papyrus: a system for high performance, distributed data mining over clusters, meta-clusters and super-clusters. *International KDD'98 Conference* (1998) pp. 37-43.
6. Kargupta H., Park B., Hershberger, D. and Johnson, E., Collective data mining: a new perspective toward distributed data mining. In HILLOL KARGUPTA AND PHILIP CHAN (eds.) *Advances in Distributed and Parallel Knowledge Discovery* (AAAI Press 1999).
7. Kimm H. and Ryu T.-W., A framework for distributed knowledge discovery system over heterogeneous networks using CORBA. *KDD2000 Workshop on Distributed and Parallel Knowledge Discovery* (2000).
8. Rana O.F., Walker D.W., Li M., Lynden S. and Ward M., PaDDMAS: parallel and distributed data mining application suite. *Proc. International Parallel and Distributed Processing Symposium (IPDPS/SPDP)*, (IEEE Computer Society Press, 2000) pp. 387-392.
9. Stolfo S.J., Prodromidis A.L., Tselepis S., Lee W., Fan D.W., Chan P.K., JAM: Java agents for meta-learning over distributed databases. *International KDD'97 Conference* (1997) pp. 74-81.

Quinta
Conferenza
Nazionale
di Statistica

WORKSHOP

Data warehousing in ambiente statistico

coordinatore

Federico Rajola

docente dell'Università cattolica di Milano

Roma 15, 16, 17 novembre 2000
Palazzo dei Congressi
Piazzale Kennedy - EUR

ISTAT

WORKSHOP

Data warehousing in ambiente statistico

coordinatore

Federico Rajola
docente dell'Università cattolica di Milano

SOMMARIO

F. Rajola	L'oganizzazione delle attività di data warehouse e data mining nel settore finanziario
D. Vanzanelli	Il data warehouse nelle banche e nelle istituzioni finanziarie: ambiti applicativi e approcci allo sviluppo
R. Torlone	Aspetti metodologici nella costruzione di un data warehouse statistico
W Lanzani	SAS
G. D'angiolini	Considerazioni su utilizzo di tecnologie OLAP/Data Warehousing in ambiente statistico
E. Giovannini, A. Sorce	Il data warehouse statistico, come fonte per la disseminazione dell'informazione e controllo di qualità
G. Bonello M.T. Doriguzzi, A. Dal Borgo	I servizi decisionali nella Pubblica Amministrazione piemontese

WORKSHOP

Data warehousing in ambiente statistico

L'organizzazione delle attività di data warehouse e data mining nel settore finanziario

F. Rajola

1. L'ORGANIZZAZIONE DELLE ATTIVITÀ DI DATA WAREHOUSE E DATA MINING NEL SETTORE FINANZIARIO

Federico Rajola
Università Cattolica del Sacro Cuore

1.1 Premessa

Nell'ultimo decennio la configurazione dei mercati creditizi registra la progressiva crescita dei livelli di competizione[1] dovuti alla caduta di restrizioni alla autonomia operativa delle banche, all'ingresso di operatori non bancari nel sistema dei pagamenti e al nuovo ruolo delle banche nei mercati finanziari.[2]

Questa progressiva trasformazione nella attività degli intermediari bancari ha stimolato e spinto il settore a supportare l'innovazione dei servizi offerti, attraverso il ricorso, sempre più coordinato e sistematico, alle tecnologie dell'informazione.

Le aziende di credito, in questo contesto evolutivo, sollecitate dai mutamenti ambientali, di mercato e legislativi intercorsi e dalle potenzialità offerte dalla tecnologia, hanno risposto agli stimoli adottando e accrescendo il proprio patrimonio tecnologico.

A ciò si aggiunga che il ricorso delle banche alle tecnologie dell'informazione è determinato dalla natura stessa dei servizi bancari, a base prevalentemente informativa.[3]

[1] Ci si riferisce in questo contesto alle innovazioni introdotte nell'ordinamento prima con il recepimento delle direttive comunitarie di coordinamento bancario, poi con il nuovo Testo Unico delle leggi in materia bancaria e creditizia.

[2] BANCA D'ITALIA, *Tematiche aziendali: la tecnologia dell'informazione e la Banca d'Italia,* II ed. nov. 1996

[3] PORTER M. E. - MILLAR V. E., "How Information Gives You Competitive Advantages", Harvard Business Review, July - August, 1985.

Infatti, il ricorso all'IT da parte delle banche risulta fondamentale non soltanto per l'automazione di elevati volumi di dati, per l'esecuzione di transazioni on-line e per la corretta esecuzione dei processi ma anche per il miglioramento del processo decisionale.

Sebbene l'innovazione tecnologica abbia, già da tempo, messo a disposizione delle banche funzionalità in grado di contribuire a migliorare i processi produttivi, non si può dire che si sia verificata una corrispondente crescita dei sistemi in grado di migliorare il processo decisionale. In altri termini, all'imponente uso dell'IT per l'automazione delle attività operative si contrappone un limitato utilizzo di efficaci sistemi di supporto alle decisioni.

Le banche e le istituzioni finanziarie da decenni utilizzano sistemi informativi per l'automazione di attività strettamente operative o transazionali. Il disegno delle architetture dei sistemi e la struttura degli archivi sono quasi sempre stati orientati al miglioramento dell'efficienza dei processi. Di conseguenza le modalità di memorizzazione e l'accessibilità ai dati hanno unicamente supportato l'attività transazionale, limitando le possibilità di riutilizzo degli stessi per fini decisionali.

Soltanto in questi ultimi anni si sta cercando di far uso dei dati disponibili per produrre informazioni necessarie ad accrescere il rigore, la robustezza e l'affidabilità nella gestione dei processi decisionali.[4]

I sistemi tradizionali hanno dimostrato l'incapacità di rispondere ad esigenze decisionali. Per far fronte alla inidoneità dell'informatica tradizionale ai suddetti fini, sono emersi nel tempo nuovi approcci. Tra essi i sistemi di *business intelligence*, ed in particolare il data warehouse e il data mining, si contraddistinguono nel contribuire al miglioramento dei processi decisionali.[5]

L'obiettivo del presente lavoro è quello di fornire un quadro di riferimento concettuale nella prospettiva dell'integrazione tra il sistema informativo tradizionale orientato alle transazioni e il sistema informativo orientato alle decisioni, nonché l'occasione per evidenziare le potenzialità del data warehouse e del data mining nel settore bancario.

Porter e Millar nell' articolo, *How Information Gives You Competitive Advantages,* attraverso la proposizione di una matrice d'intensità informativa classificano differenti settori industriali che fanno uso di tecnologie informative. Suddetta matrice consente di analizzare le esigenze di intensità informativa dei processi e di contenuti informativi del prodotto/servizio che contraddistinguono l'attività di differenti settori industriali. I due autori si soffermano a valutare le esistenti e potenziali intensità informative dei prodotti e dei processi, asserendo che queste consentono di determinare il ruolo e la dimensione dell'IT nei differenti settori aziendali. Attraverso l'analisi proposta, Porter e Millar dimostrano come l'IT sia una variabile imprescindibile dell'attività svolta dalla azienda banca. Per far ciò essi si soffermano a studiare le necessità ed i contenuti informativi propri del settore bancario che, per il perseguimento delle proprie attività, ricorrono all'IT in modo sistematico e consistente. Porter e Millar concludono, quindi, che gli elevati contenuti informativi dei prodotti/servizi e le elevate intensità informative dei processi fanno sì che il settore bancario sia uno dei settori che maggiormente si contraddistinguono per un imponente ricorso alle tecnologie dell'informazione.

Sul tema si veda anche ROSSIGNOLI C., *Organizzazione e sistemi informativi,* FrancoAngeli, Milano, 1997.

[4] Si veda a tale proposito: CIPA/ABI, *Rilevazione dello stato dell'automazione del sistema creditizio,* 1997.

[5] Si noti che già dagli anni '50 i sistemi di intelligenza artificiale, che attualmente sono ricompresi tra i sistemi di data mining, erano una tecnologia disponibile. Tali sistemi tuttavia hanno stentato ad affermarsi in parte per una naturale resistenza al cambiamento del management aziendale, in parte per gli elevati costi di implementazione ed in parte anche per impossibilità di natura tecnica quali la difficoltà di accedere e di riutilizzare i dati aziendali.

1.2 Il data warehouse

L'informatica transazionale supporta le attività *mission critical* di un'azienda. Con essa tuttavia non è possibile elaborare dati con fini differenti rispetto alla esecuzione delle transazioni. Manca, nella maggior parte dei casi, una piattaforma integrata con il sistema informativo che sia in grado di orientare i processi di elaborazione per produrre informazioni.

Nel sistema bancario l'esigenza di fornire informazioni integrate, in un linguaggio facilmente comprensibile e non di sistema, di coordinare le attività dell'azienda presentando dati che attraversano confini tecnici ed organizzativi e di rendere realmente disponibili ed utilizzabili dati storici per l'analisi degli eventi emerge sempre più prepotentemente in un mercato divenuto ormai globale e sempre più soggetto a spinte competitive.

Sotto lo stimolo del mercato, che spinge a sviluppare prodotti e servizi innovativi, a promuovere la diffusione attraverso avanzati sistemi di marketing e a misurare la redditività, si avverte una crescente attenzione delle banche ad un maggiore utilizzo delle tecnologie dell'innovazione a supporto delle strategie d'impresa[6].

Si noti, in via preliminare, che le aree applicative relative alla pianificazione, al controllo e più in generale ai sistemi informativi direzionali risultano ancora deboli, data l'impostazione piuttosto tradizionale dei vertici bancari.

Nonostante ciò, da diversi anni, l'obiettivo dei responsabili di tecnologie informatiche è quello di integrare le *system legacy application* con sistemi in grado di supportare l'attività decisionale.

Il tal senso, nonostante in passato siano stati adottati differenti approcci[7], si è compreso che le sperimentazioni intraprese non conducevano a risultati soddisfacenti e, soprattutto, non erano in grado di apportare benefici tangibili nei tempi previsti[8].

In questi ultimi anni si sta affermando un nuovo tipo di approccio: il data warehouse (DWH). Tale approccio si sta rivelando un'eccellente via per spostare il confine dell'elaborazione operazionale negli ambienti decisionali.

6 BANCA D'ITALIA, *op. cit.*, 1996.
Su tale argomento nel 1994 è stata svolta un'indagine della Nomos-Ricerca, per conto dell'ABI, che ha coinvolto 40 banche di differenti dimensioni (che alla data costituivano il 50% del totale dell'attivo del sistema bancario). Tale ricerca ha esplorato cinque differenti aree: il controllo sui diversi settori dell'area finanza, il controllo di gestione, il sistema informativo direzionale, la pianificazione strategica e l'ALM. Il risultato della ricerca ha evidenziato che per le prime due aree i sistemi informatici sono abbastanza diffusi ma non sufficientemente estesi, che la terza e la quarta area si caratterizzano per un limitato uso di strumenti informatici, molte attività vengono svolte a mano, e che la quinta area, infine, si caratterizza per un limitatissimo uso di sistemi informatici, esistono, infatti, soltanto poche realtà che fanno uso di tecnologie innovative.

7 Si pensi ad esempio al *data modeling*, all'*information engineering* e ad altri approcci di analisi dei dati.

8 DE MARCO M.,"Ruolo, prospettive e applicazioni dell'intelligenza artificiale nel mondo finanziario", in *Volume di Economia e Finanza Aziendale,* Scritti in onore di Edoardo Ardemani, Giuffrè, 1997.

Nonostante il termine DWH sia diventato molto di moda negli ultimi anni ed un gran numero di aziende stia implementando o per implementare sistemi di DWH, non esiste una unanime definizione di*magazzino di dati.*

Secondo alcuni autori il DWH è semplicemente un sinonimo di database fisico (relazionale o multidimensionale) che contiene dati; secondo altri, il DWH può essere definito come un ambiente con strutture dati finalizzate al supporto delle decisioni, fisicamente separato dai sistemi operazionali. Entrambe le definizioni, tuttavia, sembrano abbastanza limitanti e non in grado di spiegare a fondo il concetto.

Inmon, che per primo ha parlato esplicitamente di data warehouse, invece, lo definisce come una raccolta di dati **integrata**, ***subject oriented, time variant*** e ***non-volatile*** di supporto ai processi decisionali[9] Quindi, l'integrazione dei dati di un DWH costituisce una delle premesse necessarie che ne consentono una progettazione adeguata e che lo distinguono da ogni altro sistema di supporto alle decisioni.

Secondo Inmon la raccolta di dati è:

- **integrata**

  requisito fondamentale di un data warehouse è l'integrazione della raccolta dati. Nel data warehouse confluiscono dati provenienti da più sistemi transazionali e da fonti esterne.

  L'obiettivo dell'integrazione può essere raggiunto percorrendo differenti strade: mediante l'utilizzo di metodi di codifica uniformi, mediante il perseguimento di una omogeneità semantica di tutte le variabili, mediante l'utilizzo delle stesse unità di misura (p.e. giorno-mese-anno);

- ***subject oriented****:*

  perché il DWH è orientato a temi specifici dell'azienda (p.e. clienti, prodotti, etc.) piuttosto che alle applicazioni o alle funzioni (quali ad esempio in un contesto bancario ad applicazioni transazionali). In un DWH i dati vengono archiviati in modo che possano essere facilmente letti o elaborati dagli utenti. L'obiettivo, quindi, non è più quello di minimizzare la ridondanza mediante la normalizzazione ma quello di fornire dati che abbiano una struttura in grado di favorire la produzione di informazioni.

  Si passa dalla progettazione per funzioni alla modellazione dei dati al fine di consentire una visione multidimensionale degli stessi;

- ***time variant***:

  i dati archiviati all'interno di un DWH hanno un orizzonte temporale molto più esteso rispetto a quelli archiviati in un sistema operazionale. Nel DWH sono contenute una serie di informazioni relative alle aree d'interesse che colgono la situazione relativa ad un determinato fenomeno in un determinato intervallo temporale piuttosto esteso. Ciò, tuttavia, comporta che i dati contenuti in un DWH

[9] INMOM W.H., *Building the data warehouse - 2nd ed.,* John Wiley & Sons, New York, 1996, pag. 29 e seg.

sono aggiornati fino ad una certa data, che nella maggior parte dei casi, è antecedente a quella in cui l'utente interroga il sistema.
Situazione del tutto differente, al contrario, si manifesta in un transazionale in cui i dati corrispondono sempre ad una situazione costantemente aggiornata che tuttavia non fornisce un quadro storico del fenomeno analizzato;

- ***non-volatile*:**
  tale caratteristica indica la non modificabilità dei dati contenuti nel DWH che consente accessi in sola lettura. Comporta, inoltre, una maggiore semplicità di progettazione del database rispetto a quella di un database relazionale che supporta una applicazione transazionale. In tale contesto non si fronteggiano le possibili anomalie dovute agli aggiornamenti e tanto meno si ricorre a strumenti complessi per gestire l'integrità referenziale o per bloccare record a cui possono accedere altri utenti in fase di aggiornamento.

Il data warehouse, quindi, descrive il processo di acquisizione, trasformazione e distribuzione di informazioni presenti all'interno o all'esterno[10] delle aziende come supporto ai *decision maker*.

Esso si differenzia, però, in modo sostanziale dai normali sistemi gestionali che, al contrario, hanno il compito di automatizzare le operazioni di *routine*. In tabella 1 sono riportate le differenze fondamentali tra sistema transazionale ed un DWH.

[10] Ci si riferisce in questo caso ai dati contenuti negli archivi aziendali e ai dati provenienti dagli *information provider* o da internet o in generale dal sistema bancario.

Tabella 1

Differenze data warehouse e sistema transazionale

	DATA WAREHOUSE	SISTEMA TRANSAZIONALE
Utilizzo	*query-intensive*	*transaction-intensive*
Utenti	non molto numerosi	abbastanza numerosi
Storicità dei dati	dati storici e attuali	dati attuali
Integrazione	dati integrati per soggetto	dati aggregati per limitate attività o per processo
qualità dei dati	qualità intesa in termini di consistenza del dato	qualità intesa in termini di integrità dei dati
database	dati aggiornati ad intervalli predefiniti, perciò *non-volatile*	dati aggiornati continuamente e quindi *volatile*
modello dei dati	dati in genere denormalizzati; il modello è conforme alle dimensioni di un soggetto	dati normalizzati; il modello è conforme alle esigenze che derivano dalle transazioni da supportare
dominio applicativo	i progetti di data warehouse forniscono un'infrastruttura di supporto ai sistemi di supporto alle decisioni con caratteristiche di scalabilità, di ampliamento e flessibilità	i sistemi transazionali sono definiti per un limitato dominio applicativo che si riferisce a una specifica applicazione
sviluppo	i criteri di sviluppo rispondono a principi evolutivi e iterativi	sistemi OLTP[11] sviluppati seguendo i requisiti del sistema esplicitati dagli utenti e il metodo di sviluppo a cascata
Sponsorship	un progetto di data warehouse richiede una forte sponsorizzazione a causa dell'ampiezza organizzativa dello stesso	i sistemi operazionali tendono ad essere sponsorizzati seguendo un chiaro processo che consente di individuare il responsabile che individua a sua volta anche le gerarchie organizzative

Fonte: Kelly S., *Data Warehousing in Action, 1997*[12]

[11] *On Line Transaction Processing*. Il termine si riferisce agli *operational systems* che sono aggiornati on line.

[12] KELLY S., *Data Warehousing in Action,* John Wiley & Sons, New York, 1997, pag. 9

Inoltre, si noti che la definizione di Inmon sopra riportata introduce un concetto di assoluta indifferenza rispetto alle caratteristiche architetturali dei sistemi transazionali e alla dislocazione fisica dei dati nei diversi database.

Se il *focus* viene posto sulla capacità di supportare il processo decisionale, il data warehouse può essere costruito secondo modalità differenti che vanno da una logica completamente accentrata (un singolo database a cui accedono tutti gli utenti), a una logica completamente distribuita (ad esempio producendo copie di dati in ambienti tecnici e funzionali eterogenei).

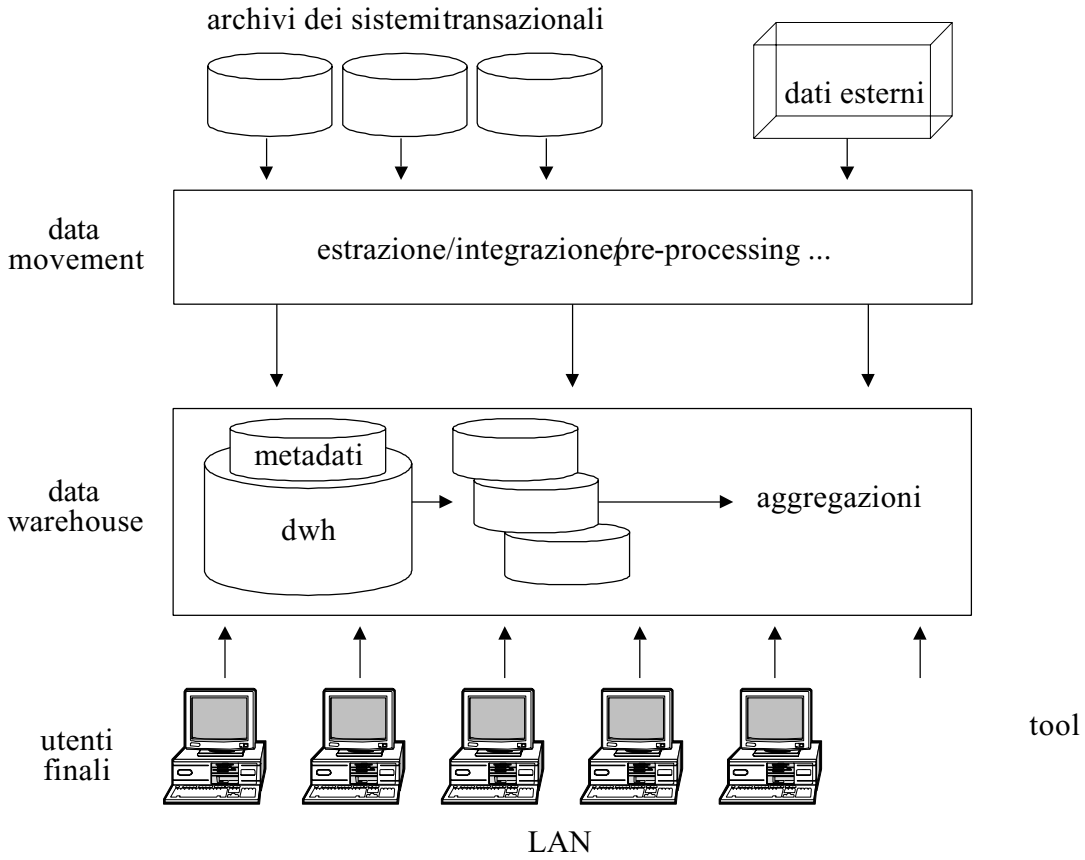
1.2.1 Le componenti e l'architettura di un DWH

In figura 1 viene riportata, a titolo esemplificativo, l'architettura generale di un data warehouse. Le principali componenti dell'architettura sono:

1. i dati provenienti dai sistemi transazionali;
2. i *data movement*
3. il data warehouse
4. i metadati
5. l'utente finale

Figura 1

L'architettura del data warehouse



I dati provenienti dai sistemi transazionali

Sono quell'insieme di dati elaborati dai sistemi transazionali della banca. Essi possono essere contenuti all'interno del medesimo database o provenire da database differenti o anche esterni all'azienda. Spesso l'architettura di un data warehouse prevede l'integrazione dei dati interni con dati esterni. L'utilizzo di questi ultimi consente di arricchire il patrimonio informativo[13].

Il data movement

Tale componente è responsabile dell'estrazione dei dati dai sistemi transazionali, dell'integrazione tra dati aziendali e dati esterni, del pre-processing dei dati, del controllo della

[13] Un esempio di dati esterni può essere quello relativo a dati di una determinata zona geografica oppure a dati provenienti dalla Centrale Rischi o a dati provenienti dagli *information provider* quali Reuters o Bloomberg, etc.

consistenza dei dati, della conversione della struttura dei dati, e dell'aggiornamento dei dizionari dati.

Il data warehouse

Dopo aver estratto i dati dagli archivi transazionali essi vengono memorizzati all'interno del data warehouse. Nel data warehouse l'accesso ai dati è consentito in sola lettura. I dati contenuti nel data warehouse hanno una dimensione storica e sono riferiti a soggetti di business.

Essi possono essere memorizzati in un repository centrale o in un data mart.

Il termine data mart è utilizzato per identificare un data warehouse di più piccole dimensioni che è orientato a supportare una particolare area di attività.[14] Si pensi, ad esempio, al data mart per il marketing in cui i dati filtrati dagli archivi transazionali sono memorizzati per supportare l'analisi della clientela. È possibile quindi che esistano all'interno della banca più data mart aventi finalità differenti e orientati a coprire diverse aree di business.[15] I dati contenuti nel data warehouse possono essere aggregati e sommarizzati per rispondere a specifiche necessità informative.

I metadati

I metadati costituiscono l'addizionale base informativa che arricchisce i dati contenuti nel data warehouse. Spesso essi vengono chiamati in gergo "data about data" indicandone la provenienza, l'utilizzo, il valore o la funzione del dato.

A tale proposito vengono costituiti dei veri e propri *information catalog*. Gli information catalog sono i file che contengono i metadati. Esso consente di spiegare all'utente la natura dei dati contenuti nel data warehouse, il loro significato semantico, da quali archivi essi provengono e la loro storicità.

L'utente finale

I dati contenuti nel data warehouse vengono presentati all'utente finale che dispone di un insieme di *tool* che consentono di effettuare elaborazioni per produrre informazioni appropriate. I tool a disposizione dell'utente possono essere semplici generatori di query e report, interfacce grafiche che consentono la rappresentazione dei dati o sistemi di analisi dati più complessi.

1.2.2 Cenni sul processo di implementazione di un data warehouse

Il processo di implementazione di un data warehouse è condizionato da un insieme di pre-requisiti di base. I principali fattori critici di successo risultano essere la componente organizzativa e tecnologica. La componente tecnologica riguarda lo studio:

[14] Talvolta il termine data mart può essere utilizzato per identificare una molteplicità di sistemi di supporto alle decisioni tra loro isolati ed orientati a specifici problemi di business. Talaltra può essere riferito ad applicazioni locali basate su server che vengono alimentate da un data warehouse centrale. In questo caso si parla anche di subsidiary data mart.

[15] È possibile avere un data mart per l'area crediti, per l'area finanza e così via.

1. delle piattaforme elaborative;
2. della configurazione delle postazioni di lavoro;
3. della rete locale;
4. del DBMS da utilizzare;
5. dell'ambiente *client/server*;
6. del software applicativo.

La componente organizzativa comprende la formazione del personale, l'acquisizione di personale con *skill* adeguati per la gestione e la manutenzione del data warehouse e la programmazione delle azioni da intraprendere.

La formazione degli utenti costituisce un fattore di primaria importanza. Si consideri, infatti, che spesso gli utenti finali sono abituati ad interagire con un sistema di tipo transazionale e che non possiedono conoscenze che consentono di effettuare l'analisi dei dati.

Se da un lato, infatti, la diffusione delle interfacce grafiche e di sistemi sempre più *user-friendly* ha favorito la diffusione di tool di analisi dei dati rendendo più semplici le attività formative degli utenti relative all'utilizzo degli stessi sistemi, dall'altro spesso mancano le conoscenze necessarie per intraprendere un'attività di analisi in grado di produrre risultati soddisfacenti.

È necessario, inoltre, che l'introduzione di un data warehouse sia accompagnata dalla creazione di nuove figure professionali interne all'azienda in grado di gestire e manutenere il sistema.[16]

Spesso può accadere che la gestione e la manutenzione del data warehouse sia demandata all'esterno. Tale circostanza, tuttavia, non esime l'azienda dall'avere almeno una figura professionale in grado di gestire dall'interno le relazioni con i fornitori e i possibili futuri sviluppi del sistema.[17]

Altra componente organizzativa riguarda la corretta programmazione e gestione delle azioni da intraprendere. A tale proposito è necessario che vengano definite:

1. le modalità di estrazione dei dati e di aggiornamento dei dati del data warehouse,
2. le modalità di accesso e i corrispondenti livelli di sicurezza (accesso al sistema mediante l'introduzione di password di protezione stratificate da assegnare ad ogni utente che consente di garantire una corretta riservatezza dei dati)
3. la gestione dei metadati.

1.2.3 L'organizzazione delle attività diwarehousing per il marketing bancario

Con l'introduzione di sistemi di data warehouse e di sistemi di analisi dati non strutturati, con l'accumulo di esperienze e con la disponibilità dei dati è diventato possibile anticipare le necessità

[16] DE MARCO M., *L'organizzazione dei sistemi informativi aziendali,* il Mulino, Bologna, 1992.

[17] Cfr. VIRTUANI R., *L'outsourcing nei sistemi informativi aziendali*, FrancoAngeli, Milano, 1997, pag. 141-146.

dei clienti. A tal fine è necessario che le aziende si focalizzino meno sull'aumento della quota di mercato e più sulla quota di ogni cliente.

Il database marketing consiste nell'utilizzo delle informazioni sul cliente per sviluppare e mantenere le relazioni con questo. Il marketing diretto da eventi aggiunge l'elemento cruciale del tempo. Dall'approccio "qui c'è il prodotto, vuoi comprarlo? " si passa a "qui c'è un cliente, cosa gli dovrei vendere, quando dovrei venderglielo e attraverso quale canale dovrei comunicare con lui?" La banca deve, quindi, essere in grado di offrire al cliente un prodotto/servizio idoneo a soddisfare le potenziali esigenze dello stesso identificando, al tempo stesso, anche il canale distributivo migliore.

Il successo di una campagna di marketing è fortemente condizionato dal grado di personalizzazione con cui l'operazione è svolta. Tuttavia, con l'eccezione di un numero ristretto di clienti di grandi dimensioni, questa attività non è commercialmente praticabile. É necessario definire in qualche modo regole di business che siano applicabili a più clienti ma che al tempo stesso diano ad esso l'impressione di soddisfare le sue esigenze specifiche. In altri termini, è necessario definire una segmentazione della clientela che aiuti a comprendere le dinamiche evolutive e le predisposizione di ogni singolo cliente.

Un effettivo orientamento al mercato significa[18] la definizione di un'offerta focalizzata sui bisogni dei vari segmenti di clientela, una comunicazione mirata e personalizzata, una struttura di canali diversificata, e così via. L'orientamento al cliente è fondato in primo luogo sui sistemi evoluti di organizzazione, analisi e impiego delle informazioni sulla clientela e più in generale sul mercato. Sono necessarie, a tale fine, strumenti e competenze avanzate in tema di acquisizione, analisi e impiego delle informazioni di mercato. Occorre disporre di una *marketing intelligence,* cioè di un sistema di monitoraggio costante del mercato a fini sia strategici sia operativi. Un tale sistema deve essere in grado di supportare le seguenti funzioni fondamentali, peraltro da svolgersi in stretta correlazione e coerenza tra di loro:

1. acquisizione dei dati interni ed esterni all'impresa;
2. analisi e interpretazione dei dati e dei fenomeni cui gli stessi si riferiscono;
3. utilizzo dei dati nel processo decisionale.

Attraverso la conoscenza della clientela è possibile, non soltanto, capire quali sono i target più interessanti ma anche comprendere come evolve lo scenario nel quale si opera, rendendo così più facile indirizzare l'azione commerciale di sviluppo verso specifici segmenti in maniera prevalente.[19]

Per far fronte a tali esigenze è utile far uso di un insieme di strumenti quali carte discoring, profili, eventi.

I modelli di carte discoring possono includere:

1. modelli di valore economico della relazione (ad es. alto valore, basso valore),
2. modelli di propensione all'acquisto di prodotti, con liste di eventi associati,
3. modelli di preferenza del canale di vendita.

[18] SCOTT W.G., "Presentazione" in BERTUCCI M., *Conoscere il cliente,* EDIBANK, Milano, 1997, pag. 7 e 8.

[19] CAMUFFO A. - COSTA G., *Banca & Organizzazione,* EDIBANK, Milano, 1995, pag. 245.

I profili o segmenti rappresentano gruppi di clienti che di solito si caratterizzano per analoghi comportamenti e per sufficienti elementi di diversità rispetto a clienti appartenenti ad altri gruppi. Possono essere basati sullo *status* (ad esempio, pensionati, coniugato, con figli, etc.), sull'utilizzo dei prodotti, o su una combinazione di entrambi.

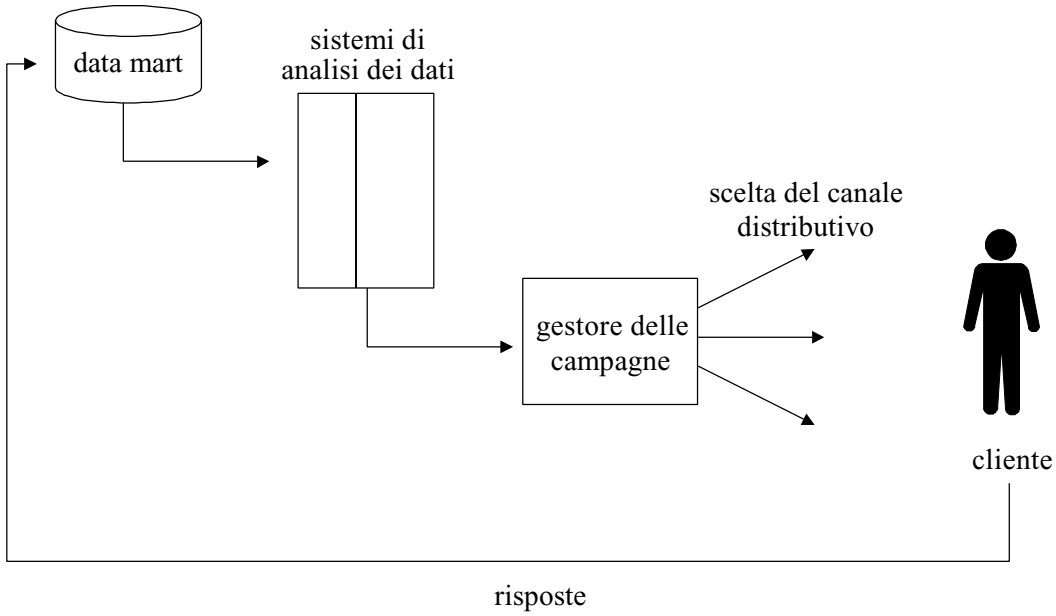
Gli eventi coincidono con i cambiamenti dei valori delle carte di scoring o di appartenenza ad un profilo, o essere definiti da eventi specifici sul database come transazioni di importo rilevante, cancellazioni di ordini permanenti, diminuzione del numero della transazioni, accrediti rilevanti con frequenza regolare. Combinando questi modelli in un gestore di campagne adeguato, si può effettuare un marketing diretto da eventi.

I componenti di una soluzione integrata di database marketing sono (figura 2):

1. il database orientato alla clientela (data mart per il marketing)
2. gli strumenti di analisi (sistemi tradizionali e sistemi di data mining)
3. il gestore delle campagne
4. i canali di distribuzione
5. gli strumenti di memorizzazione delle risposte

Figura 2

Componenti di una soluzione integrata del data base di marketing



Il database dei clienti deve contenere tutti i dati disponibili, organizzati intorno all'entità cliente, di solito in un data mart per il marketing. Benché le campagne di marketing possano essere svolte a livello famiglia o prodotto, la maggior parte delle comunicazioni sono dirette a clienti individuali, ed è per questo che il database deve essere organizzato per clienti.

Vi è una varietà di sistemi di analisi disponibili sul mercato per il *pattern recognition*, la regressione lineare, i *decision tree*, le reti neurali, la *rule induction* o gli algoritmi genetici, e in

generale tutti i sistemi di data mining. L'impiego di questi sistemi consente di ottenere un insieme di regole che individua gruppi di clienti o comportamenti.

Il gestore delle campagne integra le informazioni (ad es. la carta discoring relativa alla propensione di acquisto) e crea una serie di campagne utilizzando diverse combinazioni di dati relativi alla clientela.

I canali di distribuzione (ad es. i call center, le filiali, i promotori, la posta, internet, gli sportelli automatici), anch'essi individuabili dal gestore di campagne, sono le modalità attraverso cui i prodotti/servizi vengono offerti al cliente.

Per personalizzare le future comunicazioni è importante che vengano memorizzate le risposte del cliente all'offerta. Se da un lato è facile verificare se un cliente ha acquistato un prodotto nel periodo immediatamente successivo alla comunicazione, dall'altro risulta più difficile valutarne il comportamento se questo è avvenuto a rilevante distanza di tempo dall'offerta.

Come già detto in precedenza, uno dei punti cruciali di una fruttuosa operazione di marketing, è rappresentato dalla consistenza dei dati relativi ai clienti. Si ricordi, infatti, che l'anagrafica dei clienti proviene da archivi transazionali spesso non aggiornati. Si tenga anche presente che alla luce delle numerose operazioni di fusione ed acquisizione la complessità di gestione degli archivi anagrafici potrebbe aumentare considerevolmente. Le operazioni di M&A tuttavia possono anche rappresentare un momento di completa revisione degli stessi. È proprio nella conversione e nell'allineamento dei diversi sistemi informativi interessati alla fusione che si può prevedere un'opera di ricostruzione e "pulizia" dei dati anagrafici. Più in generale, le strutture e i sistemi a disposizione delle banche italiane per far fronte con successo alle nuove sfide competitive richiedono di essere potenziate, sia dal punto di vista strutturale, sia da quello delle competenze interne in grado di consentire il più efficace impiego possibile.[20]

Conoscere, ad esempio, la professione del cliente, il suo reddito o l'età dei figli conferisce all'azienda un reale vantaggio nel personalizzare i propri servizi. Questi dati possono essere ottenuti attraverso questionari e i moduli di richiesta dei prodotti. I clienti sono abituati a rispondere a questi tipi di domande per ogni tipo di prodotto non bancario. È importante, quindi, riconoscere il valore di questa informazione e modificare i processi organizzativi per catturare i dati e mantenerli aggiornati.

Ulteriore difficoltà è rappresentata dalla possibilità di accedere ai dati ed alle informazioni contenute negli archivi transazionali da parte degli utenti del marketing. Riscontrata ormai la suddetta difficoltà, in alcune banche è stato creato un warehouse aziendale, utilizzato come base per il database marketing. In altre, è stato creato un database di marketing separato (a volte un data mart), i cui dati sono estratti dal data warehouse aziendale o direttamente dai sistemi operazionali. Si parla in proposito di sistema informativo di marketing gestito quasi interamente dalla funzione marketing che ha previsto l'integrazione del transazionale con fonti informative esterne e con risultati delle campagne per singolo cliente. La funzione marketing può dotarsi in *staff* di una unità organizzativa, composta anche da esperti informatici, preposta allo sviluppo ed alla gestione del proprio sistema informativo dipartimentale. Tuttavia si consideri che, in molti casi, i progetti di realizzazione di un data warehouse non sono stati in grado di soddisfare le attese

[20] SCOTT W.G., *op. cit.,* 1997.

degli utenti. Nei capitoli successivi sono analizzate le principali cause che condizionano il buon fine di un progetto di warehouse.

Ai fini di una maggiore chiarezza espositiva risulta opportuno introdurre brevemente una esemplificazione del processo di analisi sulla clientela.

Da un'analisi può scaturire che un certo *cluster* di clienti risulti interessato ad un determinato prodotto. Al contrario, se il cliente ha già affermato di non essere interessato a quello specifico prodotto o di possederlo già, tale informazione assume un ruolo preminente rispetto ai risultati ottenuti nell'analisi dei dati. Un sistema in grado di fornire tali indicazioni deve quindi integrare sia i risultati dell'analisi sia le risposte fornite dal cliente in passato.

Questo breve esempio illustra che il marketing focalizzato sul cliente richiede, non soltanto, la conoscenza di quali prodotti ha il cliente ma anche la predisposizione di un ambiente che consenta di creare nel tempo un dialogo con lo stesso e che abbia memoria del passato riferita anche ai diversi canali distributivi.

È quindi utile integrare le comunicazioni di marketing con i diversi canali distributivi.

In particolare i sistemi di filiale dovrebbero consentire all'operatore di adeguare le attività di tipo informativo all'evento da gestire. L'operatore di sportello avrà pertanto bisogno di un supporto informativo multifunzionale che gli consenta la gestione diversificata e integrata delle sue applicazioni (di sportello, di consulenza, di marketing, etc.).[21] Ad esempio, per lo svolgimento delle attività di marketing in filiale è necessario che il sistema gestisca sia il messaggio di vendita sia la risposta del cliente. Uno dei vantaggi di questo canale è l'interazione immediata: il cliente visita la filiale per effettuare un'operazione ed il messaggio di marketing è consegnato come parte del processo. Anche nell'home banking il messaggio può diventare parte del processo di effettuazione della richiesta del cliente. I bancomat possono costituire un altro canale di distribuzione, attraverso la presentazione, nei tempi di attesa, di comunicazioni che si presume abbiano un qualche interesse per il singolo cliente.

Si pensi proprio alla diffusione dei canali distributivi alternativi e all'analisi dell'operatività di ogni singolo cliente. Grazie ai sistemi di*business intelligence* è possibile individuare in tempo reale caratteristiche comportamentali e propensione all'acquisto di prodotti/servizi definendo in modo dinamicopush di informazioni dirette al singolo cliente.

1.2.4 Casi di studio

Sono di seguito illustrati alcuni delle maggiori esperienze internazionali di applicazione di sistemi di businessintelligence nel settore finanziario.

Midland Bank

La Midland Bank effettua la segmentazione della clientela e il targeting utilizzando tecniche di scoring, il disegno dei test delle campagne, la valutazione dei test e delle campagne.

[21] CAMUFFO A. - COSTA G., *op. cit.*, 1995, pag. 241.

Le tecniche di targeting utilizzano i dati sui clienti e dati esterni insieme ai risultati delle campagne di promozione dei prodotti con una metodologia che individua l'importanza dei singoli dati, li combina per produrre una formula discoring, associa i valori delloscoring alla propensione di acquisto, valida la formula su nuovi dati e la utilizza su vasta scala.

Il test delle campagne valuta e ottimizza l'utilizzo di strategie alternative di offerta ed identifica gruppi di clienti per il quale una strategia è ottimale. In tal modo la campagna può utilizzare la strategia migliore per ogni cliente, anziché quella che fornisce complessivamente i risultati migliori.

First Direct

Utilizza il database marketing per vendere alla propria clientela prestiti personali, carte di credito, mutui, depositi ad alto interesse e compravendita titoli.

Grazie all'utilizzo di carte discoring e ad una strategia di test delle campagne molto accurata, First Direct ha raddoppiato i tassi di risposta ed i volumi di vendita, ha aumentato la conoscenza dei clienti, identificando quelli più redditizi, ha aumentato la propria cultura di marketing relazionale.

Bank of Scotland

Nonostante i clienti siano soddisfatti del servizio ricevuto, Bank of Scotland ha percepito il rischio che quelli più sofisticati, spesso anche i più redditizi, potessero passare alla concorrenza. Ha cercato quindi di dimostrare loro di essere in grado di fornire servizi qualitativamente più elevati rispetto alla concorrenza. Ha creato a tal fine una unità di analisi strategica nata composta da 28 persone. Vi sono analisti di mercato, pianificatori, ricercatori, statistici e modellatori di dati. Le quattro aree in cui l'unità stessa opera sono quelle relative alla qualità dei dati, al database marketing, al marketing relazionale e alla ricerca di mercato.

Il database di marketing si basa su un warehouse aziendale. È utilizzato dal gruppo di marketing relazionale che produce profili, campioni e target ed analizza l'efficacia della campagna. Il gruppo di qualità dei dati assicura che i dati siano corretti ed aggiornati.

Nel 1992 è iniziato un progetto pilota su 6 filiali utilizzando un*package* disponibile sul mercato; nel 1994 è stato creato un database di marketing completo, copia del database clienti con alcuni dati aggiuntivi; successivamente, il package è stato esteso a tutta la banca.

Le filiali individuano i clienti compresi in una campagna e possono offrire prodotti e immettere le risposte dei clienti. L'analisi statistica viene effettuata sul database di marketing. È in corso di costituzione un data mart di marketing con tutti i dati necessari per l'analisi statistica. Viene anche utilizzato un sistema informativo geografico per rappresentare i risultati delle campagne nelle varie filiali.

A breve la banca inizierà ad effettuare campagne guidate da eventi; sta anche sviluppando modelli che prevedono qual è il prodotto che il cliente comprerà per primo.

I benefici chiave del package sono il coordinamento e il controllo delle campagne, la capacità di gestire grandi volumi di dati, la gestione degli eventi, la definizione di*cluster* di mercato complessi, la gestione di gruppi di controllo, le capacità di analisi, la storia delle campagne effettuate su ogni cliente.

Canadian Imperial Bank of Commerce

Ha convertito l'intero database transazionale in un data warehouse creando una base informativa per un'efficiente gestione del rischio di credito e del rischio di mercato.

Il sistema ha l'obiettivo di integrare a livello mondiale i dati e di standardizzare le attività sui diversi mercati.

Il principale obiettivo di business raggiunto riguarda il miglioramento della gestione del rischio attraverso un meno rischioso posizionamento della banca sui mercati finanziari.

Wells Fargo Bank

Ha sviluppato un data warehouse centralizzato composto da più data mart. Il sistema è alimentato in automatico dal sistema transazionale. Le aree applicative coperte dal data warehouse comprendono il marketing, la gestione dei clienti, e d*operational reporting*. Il data warehouse è stato arricchito e completato nel tempo integrando i dati contenuti nei sistemi transazionali di banche acquisite da Wells Fargo. Ad esempio, con l'acquisizione di *1st Interstate* è partito un progetto per la razionalizzazione della rete di vendita. Il primo passo fatto, oltre all'integrazione dell'intero sistema informativo, è stato quello di acquisire i dati di *1st Interstate* per alimentare il data warehouse centralizzato. La realizzazione di un unico data warehouse ha consentito di effettuare un'analisi volta ad individuare le caratteristiche di ogni sportello. L'analisi ha consentito di ottenere, per ogni sportello, la profittabilità dello stesso, i clienti che potenzialmente potevano essere persi e la profittabilità per singolo cliente. In tal modo la banca è stata in grado di valutare i costi/benefici collegati alla chiusura di ogni sportello ottenendo un miglioramento del processo di razionalizzazione della rete di vendita.

Tale sistema è stato poi utilizzato per valutare la profittabilità di tutti i clienti della banca individuando anche i clienti potenzialmente migliori.

Capital One

Capital One ha destinato più di cinquanta persone alla realizzazione del data warehouse. È un data warehouse di circa due terabyte di dati che comprende circa sette milioni di conti. Gli obiettivi di tale sistema sono diretti a identificare nuovi clienti e la realizzazione di nuovi prodotti da offrire ai cluster di clientela più profittevoli. Ora Capital One, attraverso l'analisi dei dati contenuti nel data warehouse, è in grado di lanciare campagne promozionali e verificare in poche settimane la redemption di queste per mercati, aree geografiche e segmenti di clientela a cui la campagna stessa è rivolta.

Capital One è, inoltre, in grado di effettuare politiche di *mass customisation* nell'offerta dei prodotti *retail* e *wholesale.*

Grazie a tale iniziativa, *Capital One* è considerata una delle banche più innovative e flessibili del nord America.

First National Bank of Chicago

La *First National Bank of Chicago,* una delle maggiori banche statunitensi, ha realizzato un progetto di data warehouse le cui premesse hanno preso spunto dall'interesse mostrato dall'area finanza della banca.

L'obiettivo è stato quello di realizzare un sistema che consentisse di ottenere la razionalizzazione dei costi di gestione attraverso una più esatta pianificazione delle azioni future e delle attività di reporting. Il sistema, le cui dimensioni arrivano quasi a due terabyte, è in grado di offrire le proprie funzionalità a quattrocento utenti.

Il sistema è in grado di rispondere a quattro distinti requisiti definiti dagli utenti:

1. arricchimento degli archivi con metadati in grado di spiegare i contenuti delle tabelle e i dati;
2. disponibilità dei dati*on-demand*;
3. *electronic delivery* e accessibilità;
4. standardizzazione dei report.

I dati confluiti nel data warehouse, necessari a rispondere ai requisiti utente, provengono da otto differenti *legacy system.* L'integrazione dei dati è stata perseguita attraverso la definizione di dieci caratteristiche a cui il sistema doveva rispondere:

1. elementi chiave comuni
2. coerente definizione dei dati
3. razionale trasformazione dei dati
4. sincronizzazione dei tempi di acquisizione
5. possibilità di navigare tra i dati mediante l'utilizzo di metadati
6. *electronic information delivery*
7. aggregazione automatica dei dati
8. sistema basato su architettura client/server - mainframe
9. architettura flessibile
10. fonte centralizzata dei dettagli finanziari

Nella realizzazione del progetto, First Chicago ha posto particolare attenzione all'adeguamento degli assetti organizzativi, alla formazione del personale coinvolto ed ai problemi derivanti dall'introduzione del nuovo sistema.

First USA

First USA è il terzo maggiore emittente di carte di credito di Mastercard e VISA negli Stati Uniti. Ha implementato un data warehouse di oltre due terabyte. Il maggior *business driver* di First USA è stato quello di ampliare la propria quota di mercato. First USA ha, infatti, scelto come strategia quella di identificare i clienti potenzialmente più profittevoli e di proporre nuovi prodotti e servizi ai cluster ad alto valore.

1.3 Cenni sul processo di scoperta della conoscenza e sul data mining nel settore finanziario

Le transazioni finanziarie richiedono l'archiviazione e l'elaborazione di grandi moli di dati. I processi di scoperta della conoscenza consentono di effettuare l'analisi di tali dati per individuare complessi modelli di comportamento e caratteristiche delle variabili presenti negli archivi. I processi di scoperta della conoscenza e i sistemi di data mining possono essere utilizzati per un'ampia gamma di applicazioni finanziarie.

Le più diffuse applicazioni che utilizzano sistemi basati su processi di scoperta della conoscenza sono orientate all'attività di marketing. Tuttavia tali sistemi sono impiegati con successo anche in altri domini applicativi. Essi vengono utilizzati per l'analisi e la selezione di strumenti finanziari, col fine di ridurre il rischio e/o di massimizzare l'investimento, per la riduzione delle frodi, col fine di identificare comportamenti fraudolenti nell'utilizzo delle carte di credito in caso di furto o contraffazione, per la minimizzazione del rischio di credito e la per valutazione della qualità del cliente.[22]

Nel seguito del paragrafo sono esposti sinteticamente i punti fondamentali del processo di scoperta della conoscenza e dei sistemi di data mining.

1.3.1 Il processo di scoperta della conoscenza

I significati di KDD - *knowledge discovery in database* - e di data mining sono spesso utilizzati come sinonimi.

Il KDD descrive l'intero processo di identificazione ed estrazione di significativi e potenzialmente utili modelli di interpretazione della realtà necessari per generare conoscenza.

Il data mining, invece, si riferisce all'applicazione di algoritmi per l'estrazione di modelli o regole di comportamento presenti nei dati e alla corretta interpretazione dei risultati. Esso dunque rappresenta una delle fasi del processo di scoperta della conoscenza.[23] Si veda a tale proposito figura 3.

[22] Si vedano tra gli altri:
BAESTAENS D.E. e altri, *Neural Network Solutions for Trading in Financial Markets,* The Financial Times - Pitman Publishing, London, 1994.
DEBOECK G. J. (a cura di), *Trading on the Edge*, John Wiley & Sons, New York, 1995.
DE MARCO M., "Aspettative e realtà dell'intelligenza artificiale," in Aggiornamenti Sociali, n.11, 1988.
DE MARCO M., "Integrazione di sistemi esperti con il sistema informativo bancario," in GARDIN F. - ROSSIGNOLI C. - VATURI S., *Sistemi esperti banca e finanza*, il Mulino, Bologna, 1991.
NOTTOLA C. ROSSIGNOLI C. 1994, (a cura di), *Analisi della esperienza operativa sull'utilizzo dell'intelligenza artificiale in banca*, FrancoAngeli.
ROSSIGNOLI C., *Applicazioni di sistemi esperti e reti neurali un campo finanziario*, FrancoAngeli, Milano, 1993.

[23] FAYYAD U. M. - SHAPIRO G. (a cura di), *Advances in Knowledge Discovery and Data Mining,* AAAI - The Mit Press, Menlo Park, 1996.

Il processo di scoperta della conoscenza è un processo iterativo ed interattivo[24] che richiede all'utente di percorrere numerose fasi per arrivare a "scoprire conoscenza".

Tale processo può essere schematicamente riassunto come segue:

1. definizione e comprensione del dominio applicativo e definizione degli obiettivi di business da realizzare;
2. creazione di un *target data set*: selezione di un *subset* di variabili e di dati o campionamento dei dati da utilizzare per il processo di scoperta della conoscenza;
3. *data cleaning* e *pre-processing:* operazioni basilari per attenuare la presenza del rumore nei dati, o degli *outlier* se necessario, selezione delle informazioni necessarie per generare il modello, decisioni sulle modalità di trattamento dei campi mancanti o incompleti, sulla definizione della dimensione storica dei dati da trattare e definizione delle modalità di aggiornamento;
4. *data reduction* e *projection*: definizione della modalità di rappresentazione dei dati collegate agli obiettivi da realizzare. Utilizzo dei metodi di trasformazione per ridurre il numero delle variabili;
5. scelta del ruolo dei sistemi di data mining per l'analisi: utilizzo dei sistemi di data mining per classificazione, regressione, clusterizzazione, etc.
6. scelta del o degli algoritmi di data mining: selezione dei metodi da utilizzare per la ricerca di pattern. Tale scelta richiede di decidere quali modelli o parametri possono essere più appropriati. Integrazione dei metodi di data mining scelti con l'intero processo di scoperta della conoscenza;[25]
7. data mining: ricerca di modelli di interesse per l'utente presentati secondo definite modalità di rappresentazione: classificazione, induzione di regole, *decision tree*, regressione, definizione di cluster etc. L'utente può interagire con il sistema per raffinare i risultati di volta in volta ottenuti;
8. interpretazione dei modelli identificati, possibile retroazione ai punti precedenti per ulteriori iterazioni;
9. consolidamento della conoscenza scoperta: integrazione della conoscenza e valutazione delle *performance* del sistema. Produzione della documentazione agli utenti finali o a terze parti interessate.

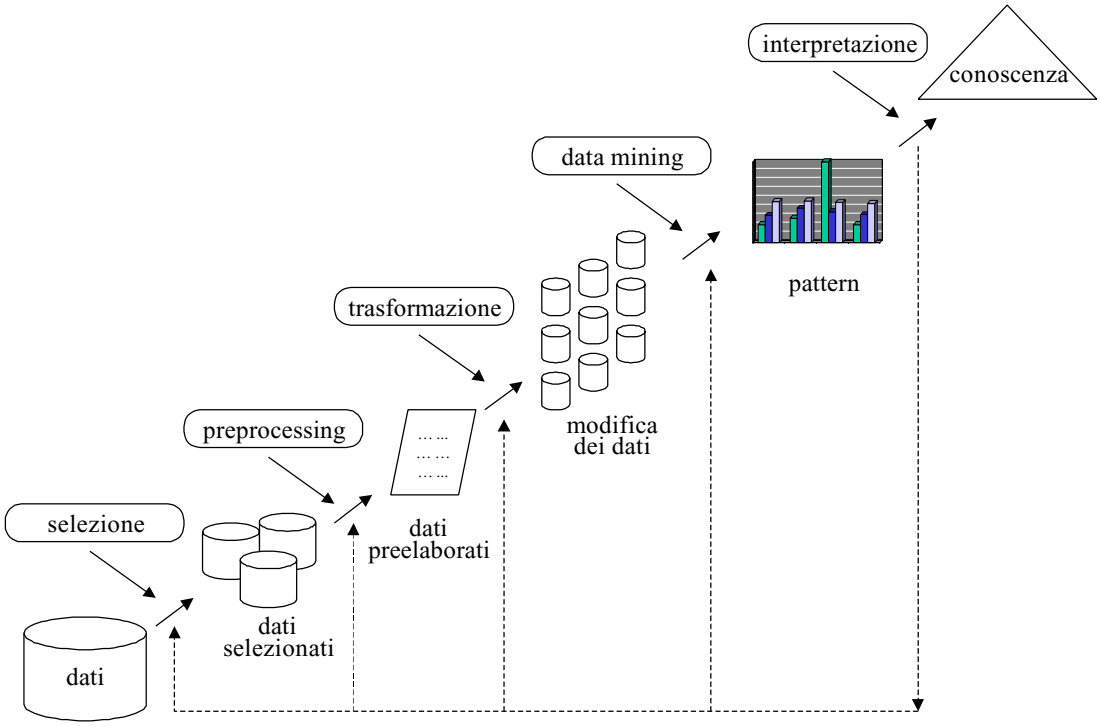
È importante notare che tale processo non si conclude con un'unica iterazione. Esso richiede di raffinare progressivamente i risultati ottenuti per pervenire alla realizzazione di un modello in grado di interpretare i fenomeni oggetto di analisi.

[24] BRANCHMAN R. J. - ANAND T., "The Process of Knowledge Discovery in Databases: A Human-Centered Approach", FAYYAD U. M. - SHAPIRO G. (a cura di), *op cit*., 1996, pag. 37 e seg.

[25] Ad esempio l'utente può essere più interessato a capire il modello che a derivare una previsione.

Figura 3

Il processo del KDD



Fonte: FAYYAD U. M. - SHAPIRO G.[26]

Il processo di scoperta della conoscenza può richiedere un significativo numero di iterazioni, come rappresentato dalla linea tratteggiata in figura 3. Esso può anche dar luogo a *loop* tra due o più fasi. In figura 3 sono rappresentati i passi fondamentali per arrivare a scoprire conoscenza. È data anche rappresentazione schematica dei possibili *loop* tra le diverse fasi.

Dopo aver sinteticamente individuato i punti fondamentali del processo di scoperta della conoscenza, si focalizza ora l'attenzione sui sistemi di data mining.

1.3.2 Il data mining

Il data mining, come già detto, è uno dei componenti del processo di scoperta della conoscenza, può essere definito come un insieme di tecniche che consentono di effettuare l'esplorazione e

[26] FAYYAD U. M. - SHAPIRO G., *op cit*., 1996, pag. 10.

l'analisi dei dati per scoprire significative regole o modelli nascosti all'interno di archivi di grandi dimensioni in modo automatico o semiautomatico.[27]

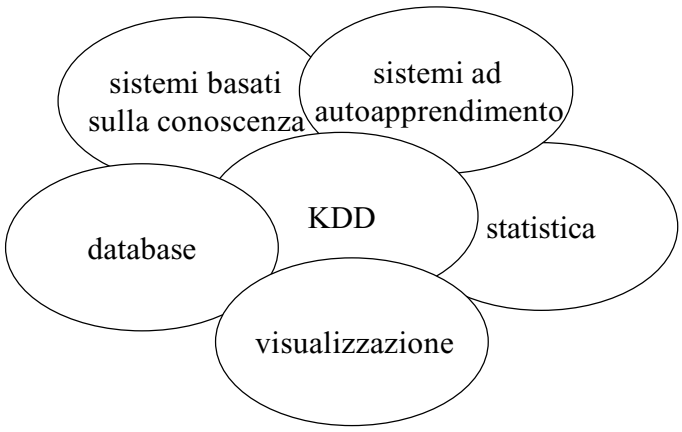
Esso è, quindi, un approccio multidisciplinare che riunisce un insieme di tecniche quali la statistica, la visualizzazione, i sistemi basati sulla conoscenza e i sistemi ad autoapprendimento che consentono di scoprire conoscenza e di tradurla in regole o modelli utili per risolvere problemi di business (figura 4).

L'obiettivo dei sistemi di data mining è di consentire, alle diverse aree applicative dell'azienda, il miglioramento dei processi conoscitivi e la riduzione dell'incertezza legata all'assunzione di decisioni.

Come per la realizzazione di qualsiasi sistema informativo, anche in questo caso è necessario partire da una approfondita analisi delle esigenze di business. Si tratta di coinvolgere, l'utente ad ogni livello di realizzazione dello stesso, perché non basta soltanto verificare i requisiti utente, ma anche validare i risultati ottenuti e determinarne i livelli di bontà. Tale aspetto non va trascurato poiché nella realizzazione dei sistemitransazionali è quasi sempre possibile individuare un processo o un algoritmo e procedere quindi con le successive fasi di automazione. Nella realizzazione di sistemi di business intelligence, e in particolare di data mining, si devono determinare proprio gli algoritmi sottostanti e quindi, dopo aver individuato i modelli di comportamento, è necessario procedere con una verifica a più *step* per un progressivo raffinamento degli stessi al fine di implementare un valido sistema.

Figura 4

L'approccio multidisciplinare del data mining



Fonte: ADRIAANS P. - ZANTINGE D.[28]

[27] BERRY M. - LINOFF G., *Data mining techniques,* Wiley & Sons, New York, 1997.

[28] ADRIAANS P. - ZANTINGE D., *Data Mining,* Addison-Wesley, Harlow, 1996.

L'utilizzo dei sistemi di data mining, per la definizione dei modelli, può avvenire secondo due differenti modalità. Esso può essere a iniziativa mista, da parte del sistema e dell'utente, o essere effettuato automaticamente da parte del sistema.

Nel primo caso l'utente è assistito da un processo intelligente che automaticamente compie l'analisi seguendo criteri pre-impostati e identificando *pattern* significativi. Si veda figura 5.

Nel secondo il sistema scopre pattern, mediante la ricerca autonoma di modelli senza alcuna richiesta impostata dall'utente. In questo caso il sistema consente di automatizzare il processo di analisi esplorativa dei dati. Tale sistema è in grado di effettuare interrogazioni autonomamente. Per scoprire informazioni esso formula un'ipotesi, effettua una interrogazione, conduce un'analisi statistica sui risultati ottenuti, visualizza i risultati e modifica le ipotesi. Questo ciclo continua finché non emerge un modello di interpretazione. In tale processo il sistema interroga il database ripetutamente, valida o rigetta i risultati parziali finché non scopre il modo per modellizzare un determinato comportamento, figura 6.

Figura 5

Interazione tra utente e sistema di data mining

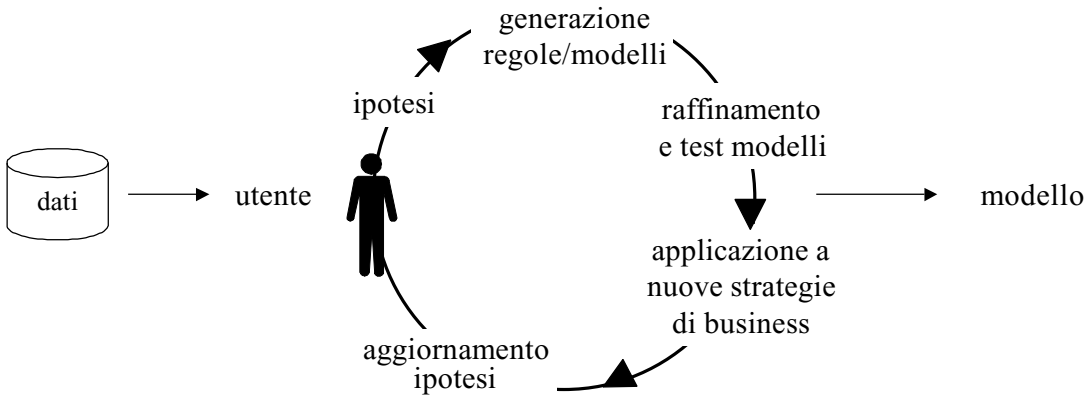
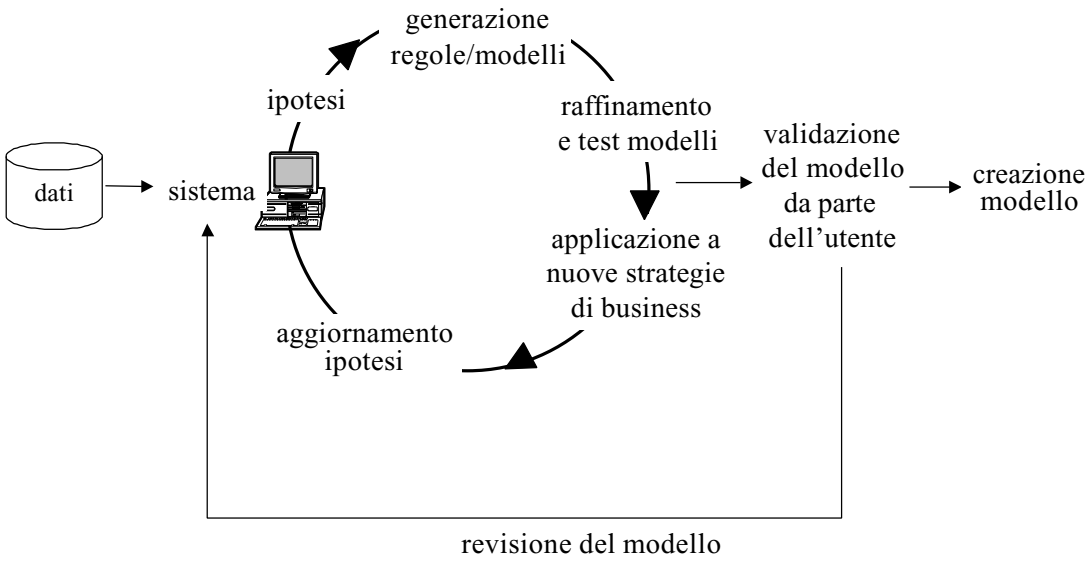


Figura 6

Generazione del modello a iniziativa del sistema



1.3.3 I principali sistemi di data mining

Come già detto in precedenza, esistono differenti tecniche di data mining. L'impiego di una o più tecniche per risolvere problemi di business è legato al dominio applicativo cui l'analisi è rivolta. Si vedano a tale proposito le tabelle 2 e 3. Sebbene siano disponibili sul mercato un'ampia gamma di algoritmi, le principali e più innovative tecniche di data mining sono[29]:

1. la visualizzazione
2. le reti neurali
3. gli algoritmi genetici
4. la *fuzzy logic*
5. i *decision tree e rule induction*

Le tecniche di visualizzazione dei dati sono metodi per la scoperta di pattern all'interno di un data set. Il data set viene di volta in volta creato dall'utente mediante l'interrogazione di un data mart o di un data warehouse. Tali tecniche possono essere utilizzate, all'inizio del processo di data mining, per verificare in modo approssimativo la consistenza e la qualità dei dati oppure per interpretare i risultati di un'analisi condotta con altre tecniche di data mining.

La visualizzazione è un sistema "trasversale" adatto per obiettivi di comprensione di fenomeni, anche dinamici, per il supporto all'analisi dei dati, da utilizzare in congiunzione con sistemi di Q&R (query and reporting), EIS (Executive Information System), analisi statistica, reti neurali.

[29] In tale contesto non verranno trattati i *query tool*, le tecniche statistiche, i *knowledge based system* e gli OLAP, Online Analytical Processing. Si rimanda in proposito il lettore alla bibliografia sul data mining fornita alla fine del capitolo.

Tabella 2

PROBLEMI	DEFINIZIONE
Classificazione	Definizione delle caratteristiche del data set
Clustering	Identificazione delle affinità che definiscono i gruppi in un data set che mostrano comportamenti simili e delle variabili che sono in comune
Sequencing	Identificazione delle correlazioni tra comportamenti all'interno di un periodo predefinito
Associazione	Identificazione delle correlazioni tra comportamenti che ricorrono nello stesso periodo
Previsione	Identificazione di trend non lineari basata su dati storici

Fonte: Meta Group, 1999

Tabella 3

ESEMPIO	TIPO DI PROBLEMA	TECNICA ADOTTABILE
Quali sono i tre principali motivi che hanno indotto il mio cliente a passare alla concorrenza?	classificazione	Reti neurali Decision tree
Quali sono le fasce di clienti a cui posso offrire nuovi prodotti/servizi?	clustering	Reti neurali Decision tree
Quali sono le probabilità che un cliente che ha aperto un c/c acquisterà anche il prodotto X in breve tempo?	sequencing	Tecniche statistiche Rule induction
Quali sono le probabilità che un cliente acquisti due prodotti completamente differenti?	associazione	Tecniche statistiche Rule induction
Quale sarà il prezzo del titolo tra un giorno/mese etc.?	previsione	Reti neurali Tecniche statistiche

Fonte: Meta Group, 1999

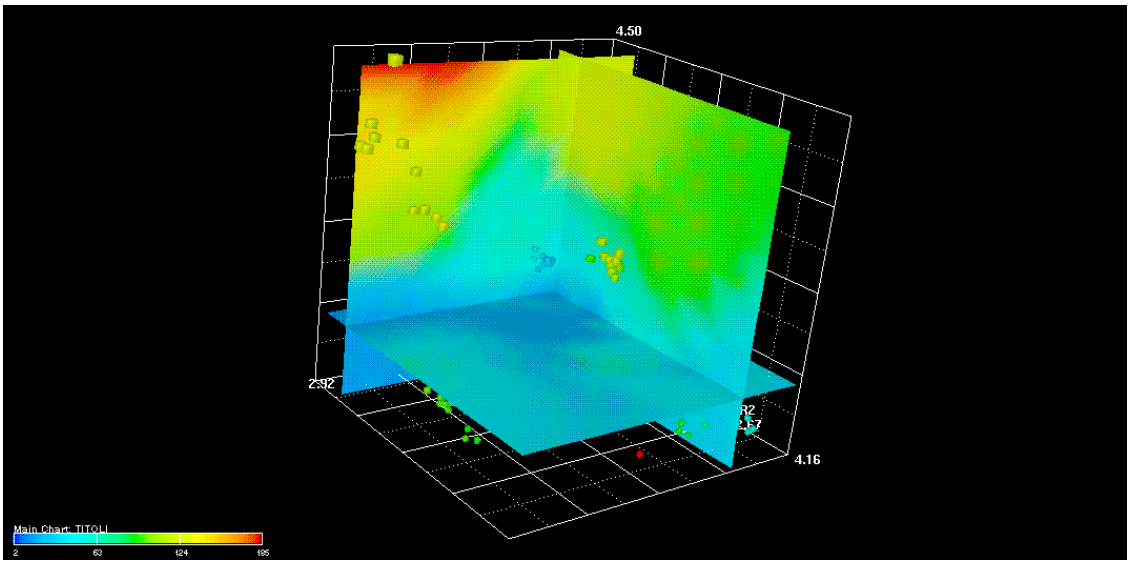
La visualizzazione consente di effettuare l'analisi di correlazioni tra variabili, l'analisi di serie temporali, l'identificazione di pattern, l'identificazione di anomalie, l'analisi di cluster e segmenti, l'analisi territoriale.[30]

[30] La visualizzazione si ispira alla *teoria della Gelstat* secondo la quale una *gelstat* è un'entità organizzata o un tutto, in cui le parti, per quanto distinguibili, sono interdipendenti; esse hanno certe caratteristiche derivate dalla loro inclusione nel tutto, e il tutto ha alcune caratteristiche che non appartengono a nessuna delle parti.

Alcuni sistemi di visualizzazione dei dati presenti sul mercato consentono di rappresentare in un unico grafico fino a dodici dimensioni.[31] A tale proposito si vedano figura 7, 8 e 9. In figura 7 è rappresentata una correlazione tra variabili differenti (finanziarie, commerciali, organizzative ecc.) di centinaia di agenzie bancarie utilizzando interpolazioni e piani di taglio. In figura 8 è rappresentata una analisi relativa alla identificazione delle anomalie operative di sportelli bancari di un grande gruppo italiano. In figura 9 è rappresentata una analisi della distribuzione dei clienti di una banca in cluster. Tale analisi permette di identificare il comportamento di classi di clienti per segmenti rispetto a determinati servizi finanziari.

Figura 7

Analisi visuale: correlazioni

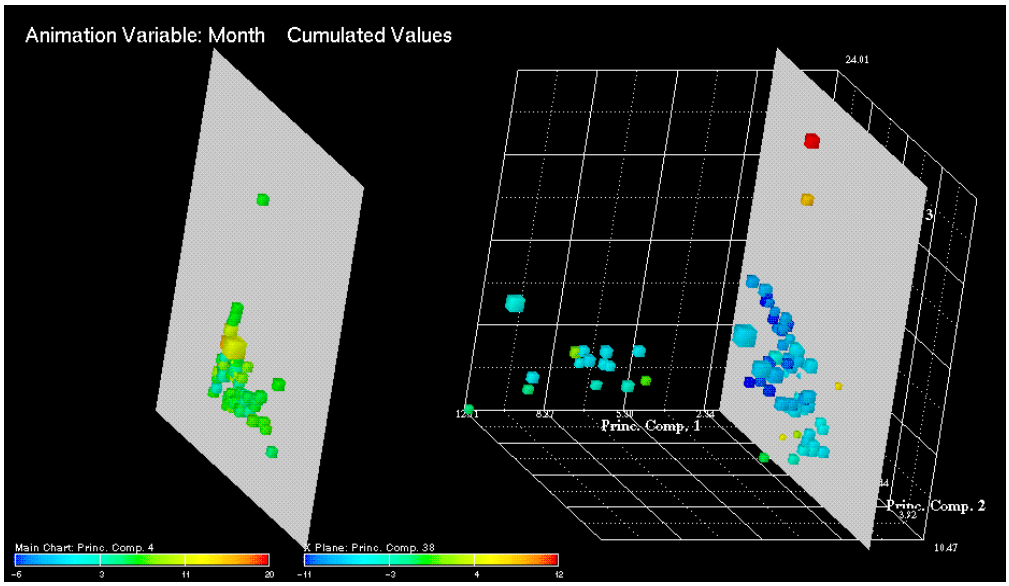


Fonte: DBInspector

Figura 8

Analisi visuale: ricerca di anomalie

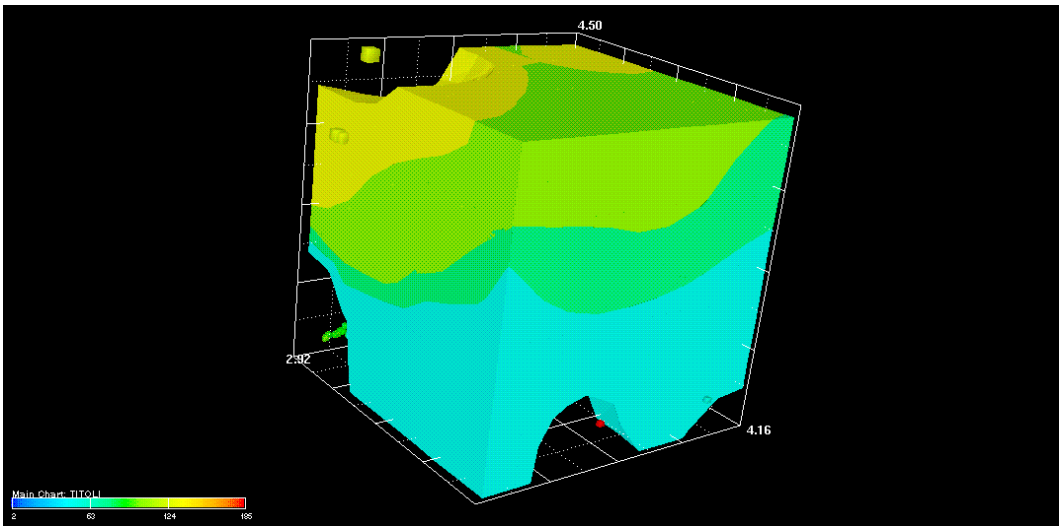
[31] Le rappresentazioni grafiche proposte sono state elaborate con un software di visualizzazione sviluppato nel progetto DBInspector finanziato dalla Comunità Europea nell'ambito del programma Esprit. I partner che hanno collaborato al progetto sono: l'Ufficio Italiano dei Cambi, il Parallel Applications Centre dell'Università di Southampton, il Centro di Tecnologie Informatiche e Finanziarie - CeTIF dell'Università Cattolica del Sacro Cuore di Milano, l'AIS e l'Università di Trento.
Esso è anche impiegato in un altro progetto finanziato dalla Comunità Europea, sempre nell'ambito del programma Esprit. In tale progetto il prodotto viene impiegato per condurre analisi comportamentali della clientela della Banca Monte dei Paschi S.p.a. Il fine è quello di individuare potenziali *cluster* di clienti interessati al risparmio gestito. I partner che collaborano al progetto sono, oltre alla Banca Monte dei Paschi di Siena, il Parallel Applications Centre dell'Università di Southampton, il Centro di Tecnologie Informatiche e Finanziarie - CeTIF dell'Università Cattolica del Sacro Cuore di Milano el'AIS.



Fonte: DBInspector

Figura 9

Analisi visuale: comportamento dei clienti rispetto a determinati prodotti della banca



Fonte: DBInspector

Le reti neurali sono tra le più diffuse tecniche di data mining disponibili sul mercato.

Esse sono in grado, mediante l'apprendimento su un insieme di dati storici chiamato *learning set,* di generare pattern e di validarli sulla base di un altro subset di dati chiamato *test set.* Operano in modo iterativo modificando di volta in volta i pattern trovati fino a determinare una soluzione ottimale. Sono in grado di effettuare previsioni su serie storiche e classificazione di eventi.

Si riscontrano numerose applicazioni in campo finanziario.[32] Esse sono sintetizzabili come segue:

1. screening per la concessione di prestiti
2. valutazione del rischio di credito su prestiti ipotecari
3. previsioni economico-finanziarie
4. diversificazione e selezione di portafoglio
5. simulazione di andamenti di mercato
6. segmentazione della clientela
7. costruzione di indici di profittabilità

[32] Per un approfondimento dell'impiego delle reti neurali in campo finanziario si vedano tra gli altri:
DEBOECK G. J., *Trading on the Edge, New York*, John Wiley & Sons, Inc., 1995.
NOTTOLA C. - ROSSIGNOLI C. (a cura di), *Intelligenza artificiale in banca: tendenze evolutive ed esperienze operative a confronto,* FrancoAngeli, Milano, 1995.
STEIN J., *Neural networks: from the chalkboard to the trading room*, Trading Techniques, New York, 1991.
TARUN K., *Foundations of Neural Networks*, Addison Wesley, New York, 1990.
TURBAN E. - TRIPPI R., *Neural networks in finance and investing: using artificial intelligence to improve real-world performance*, Probus publishing company, 1993.
UNWIN C. - COGBILL S., *Artificial Intelligence in Financial Trading* , Financial trading International, New York, 1991.

8. individuazione di frodi
9. previsione dell'andamento del mercato mobiliare
10. gestione e pianificazione delle carriere

Gli algoritmi genetici sono tecniche di ottimizzazione che traggono spunto dalla teoria darwiniana dell'evoluzione della specie. Proposti inizialmente da Holland[33] per la definizione di paradigmi computazionali, essi possono essere utilizzati per la ricerca di soluzioni a problemi di ottimizzazione e di classificazione, adottando tecniche che imitano il processo di evoluzione e selezione naturale. Mediante l'utilizzo di un algoritmo e di un *subset* di dati, sono in grado di individuare soluzioni ottimali o subottimali relative ad un determinato problema.

In campo economico e finanziario, gli algoritmi genetici sono utilizzati per l'ottimizzazione del portafoglio titoli con l'obiettivo della minimizzazione del rischio e la massimizzazione dell'investimento, per la previsione di situazioni di insolvenza, per l'individuazione delle frodi, per l'individuazione, in analisi tecnica, dei valori ottimali di medie mobili (periodo e valore dei coefficienti esponenziali).

In sintesi, caratteristica peculiare degli algoritmi genetici è quella di risolvere problemi facendo evolvere le soluzioni come avviene in natura, piuttosto che ricercare le soluzioni seguendo regole predefinite.[34]

I sistemi di *fuzzy logic* derivano dalla teoria degli insiemi sfumati e rientrano nel filone di studi della logica a valori indefiniti.[35] Lo scopo dei sistemi a logica sfumata è di consentire l'adozione, nell'ambito dei processi decisionali, di meccanismi in grado di gestire informazioni caratterizzate da elementi qualitativi o da un basso grado di decisione. Una delle funzioni principali della fuzzy logic è di facilitare la definizione degli ambiti in cui si svolgono i processi decisionali, soprattutto nella fase di interazione del programma con il mondo esterno.[36]

Per questo la fuzzy logic è spesso considerata come uno strumento flessibile attraverso il quale un sistema può, allo stesso tempo, ricevere istruzioni e dare spiegazioni all'utente sulle azioni intraprese.[37]

Un esempio di applicazione in campo finanziario può essere quello relativo alla valutazione del peso di una variabile qualitativa sull'andamento di un indice di borsa. Nell'ipotesi in cui si voglia attribuire un valore quantitativo ed oggettivo ad un evento politico che può condizionare l'indice di borsa, che tralasci qualsiasi considerazione soggettiva, si può far uso di un sistema di logica sfumata. Esso è in grado di parametrizzare una variabile qualitativa e trasformarla in quantitativa ed oggettiva.

[33] Gli algoritmi genetici sono stati progettati per imitare alcuni processi osservati durante l'evoluzione naturale degli organismi viventi. Holland pensò di incorporare le strutture di tali processi in algoritmi matematici. HOLLAND J.H., "Genetic Algorithms and the Optimal Allocations of the Trials", in SIAM Journal of Computing, n. 2, 1973.

[34] ROSSIGNOLI C., *op. cit.,* 1997, pag. 142.

[35] DUBOIS D. - PRADA H., "Fuzzy Sets and Systems: Theory and Application", Academic Press, San Diego, 1980. ZADEH L., "Making Computers Think Like People", in IEEE Spectrum, vol. 21/8, 1984, pag. 26-32

[36] ROSSIGNOLI C., *op. cit.*, 1997, pag. 143.

[37] KINGDON J. - FELDMAN K., "Intelligent Techniques Applied in Finance", in NOTTOLA C. - ROSSIGNOLI C. (a cura di), *Intelligenza artificiale in banca: tendenze evolutive ed esperienze operative a confronto,* Franco Angeli, Milano, 1995.

I *decision tree* e la *rule induction* sono modelli che operano secondo la seguente modalità: dividono il subset di dati in insiemi più piccoli, ognuno dei quali è spiegato da regole. L'output di tali sistemi consente di identificare regole e modelli in modo automatico mediante l'apprendimento sui dati storici. Una volta identificate, le regole o i modelli possono essere utilizzati per l'analisi di nuovi subset di dati. Il vantaggio è rappresentato dal fatto che l'utente è in grado di identificare e analizzare le regole individuate dal sistema; il limite è, invece, costituito dal fatto che spesso producono un gran numero di regole difficili da gestire e da interpretare.

Tabella 4

Vantaggi e limiti dei più diffusi sistemi di data mining

TECNICHE	VANTAGGI	LIMITI
Visualizzazione	L'utente è in grado di visualizzare grandi moli di dati, di scoprire relazioni e di testarle	Richiede un utente esperto in statistica e in grado di utilizzare altre tecniche di data mining
Reti neurali	Elevata capacità elaborativa con dati in cui si nascondono relazioni non lineari. In grado di lavorare con dati incompleti e "rumorosi"	Incapacità di fornire spiegazioni sui risultati ottenuti sebbene sia possibile utilizzare altri sistemi in grado di fornire un'interpretazione. I dati qualitativi devono essere convertiti in quantitativi.
Algoritmi genetici	Buona capacità previsionale utilizzando dati in cui si nascondono relazioni non lineari.	Incapacità di fornire spiegazioni sui risultati ottenuti sebbene sia possibile utilizzare altri sistemi in grado di fornire un'interpretazione. I dati qualitativi devono essere convertiti in quantitativi.
Fuzzy logic	Può classificare variabili e risultati sulla base di "vicinanza" alla soluzione desiderata	Numero limitato di fornitori e applicazioni disponibili sul mercato
Decision tree e *rule induction*	Creano regole e modelli sulla base di dati storici. Le regole e i modelli sono trasparenti all'utente e facilmente interpretabili	Richiedono un*tuning* ottimale per evitare la produzione di elevati numeri di regole difficilmente interpretabili e gestibili

Nel seguito del volume sono illustrate le metodologie di realizzazione di sistemi di business intelligence, le architetture hardware persupportarli, alcuni dei più diffusi prodotti di mercato e due casi di applicazione nel settore bancario.

Bibliografia

ADRIAANS P. - ZANTINGE D., *Data Mining,* Addison-Wesley, Harlow, 1996.

BAESTAENS D.E. e altri, *Neural Network Solutions for Trading in Financial Markets,* The Financial Times - Pitman Publishing, London, 1994.

BANCA D'ITALIA, *Tematiche aziendali: la tecnologia dell'informazione e la Banca d'Italia,* II ed. nov. 1996

BERRY M. - LINOFF G., *Data Mining Techniques,* Wiley & Sons, New York, 1997.

BRANCHMAN R. J. - ANAND T., "The Process of Knowledge Discovery in Databases: A Human-Centered Approach", FAYYAD U. M. - SHAPIRO G. (a cura di), *Advances in Knowledge Discovery and Data Mining,* AAAI - The Mit Press, Menlo Park, 1996.

CAMUFFO A. - COSTA G., *Banca & Organizzazione,* EDIBANK, Milano, 1995.

CIPA/ABI, *Rilevazione dello stato dell'automazione del sistema creditizio*, 1997.

DE MARCO M., *L'organizzazione dei sistemi informativi aziendali,* il Mulino, Bologna, 1992.

DE MARCO M., "Aspettative e realtà dell'intelligenza artificiale," in Aggiornamenti Sociali, n. 11, 1988.

DE MARCO M., "Integrazione di sistemi esperti con il sistema informativo bancario," in GARDIN F. - ROSSIGNOLI C. - VATURI S., *Sistemi esperti banca e finanza*, il Mulino, 1991.

DE MARCO M.,"Ruolo, prospettive e applicazioni dell'intelligenza artificiale nel mondo finanziario", in *Volume di Economia e Finanza Aziendale,* Scritti in onore di Edoardo Ardemani, Giuffrè, 1997.

DEBOECK G. J. (a cura di), *Trading on the Edge*, John Wiley & Sons, New York, 1995.

DUBOIS D. - PRADA H., "Fuzzy Sets and Systems: Theory and Application", Academic Press, San Diego, 1980.

FAYYAD U. M. - SHAPIRO G. (a cura di), *Advances in Knowledge Discovery and Data Mining,* AAAI - The Mit Press, Menlo Park, 1996.

HOLLAND J.H., "Genetic Algorithms and the Optimal Allocations of the Trials", in SIAM Journal of Computing, n. 2, 1973.

INMOM W.H., *Building the Data Warehouse - 2nd ed.,* John Wiley & Sons, New York, 1996.

KELLY S., *Data Warehousing in Action,* John Wiley & Sons, New York, 1997.

KINGDON J. - FELDMAN K., "Intelligent Techniques Applied in Finance", in NOTTOLA C. - ROSSIGNOLI C. (a cura di), *Intelligenza artificiale in banca: tendenze evolutive ed esperienze operative a confronto,* FrancoAngeli, Milano, 1995.

NOTTOLA C. - ROSSIGNOLI C. (a cura di), *Intelligenza artificiale in banca: tendenze evolutive ed esperienze operative a confronto,* FrancoAngeli, Milano, 1995.

NOTTOLA C. ROSSIGNOLI C., (a cura di), *Analisi della esperienza operativa sull'utilizzo dell'intelligenza artificiale in banca*, FrancoAngeli, Milano, 1994.

PORTER M. E. - MILLAR V. E., "How Information Gives You Competitive Advantages", Harvard Business Review, July - August, 1985.

ROSSIGNOLI C., *Applicazioni di sistemi esperti e reti neurali un campo finanziario*, FrancoAngeli, Milano, 1993.

ROSSIGNOLI C., *Organizzazione e sistemi informativi,* FrancoAngeli, Milano, 1997.

SCOTT W.G., "Presentazione" in BERTUCCI M., *Conoscere il cliente,* EDIBANK, Milano, 1997.

STEIN J., *Neural networks: from the Chalkboard to the Trading Room*, Trading Techniques, New York, 1991.

TARUN K., *Foundations of neural networks*, Addison Wesley, New York, 1990.

TURBAN E. - TRIPPI R., *Neural networks in Finance and Investing: Using Artificial Intelligence to Improve Real-World Performance*, Probus publishing company, New York, 1993.

UNWIN C. - COGBILL S., *Artificial Intelligence in Financial Trading*, Financial trading International, New York, 1991.

VIRTUANI R., *L'outsourcing nei sistemi informativi aziendali*, FrancoAngeli, Milano, 1997.

ZADEH L., "Making Computers Think Like People", in IEEE Spectrum, vol. 21/8, 1984.

WORKSHOP

Data warehousing in ambiente statistico

Il data warehouse nelle banche e nelle istituzioni finanziarie: ambiti applicativi e approcci allo sviluppo

D. Vanzanelli

1. IL DATA WAREHOUSE NELLE BANCHE E NELLE ISTITUZIONI FINANZIARIE: AMBITI APPLICATIVI E APPROCCI ALLO SVILUPPO

1.1. Perchè il data warehouse?

Immaginiamo per un attimo di trovarci in un immenso magazzino.

In questo magazzino ci sono sicuramente tantissimi oggetti interessanti, il problema è che alcuni sono interessanti per alcune persone, altri sono interessanti per altre ed è evidente che la loro disposizione non è molto incoraggiante per qualcuno che deve cercare qualcosa senza perdere troppo tempo.

Immaginiamo che:

- questo sia il magazzino di un negozio di articoli sportivi
- in questo magazzino ci siano TUTTI gli attrezzi di TUTTE le marche per TUTTI gli sport
- il commesso non ci sia o non abbia voglia di aiutarci
- il magazzino sia organizzato in base a criteri a noi totalmente ignoti e comunque incomprensibili
- noi siamo appassionati giocatori di tennis

- noi stiamo cercando un tubo di palle per giocare sull'erba

Nel cercare ciò che ci serve (le palle da erba) probabilmente troveremo diverso materiale attinente (scarpe bianche da terra rossa, polo bianche, palline da tartan, racchette in titanio, asciugamani firmati Venus Williams) scarteremo con disprezzo altri oggetti come scarpe da basket alte fino al ginocchio, mazze da golf, ginocchiere da volley e spade da kendo. Tutti questi oggetti sarebbero invece l'obiettivo primario di altre persone che scarterebbero con altrettanto disprezzo ciò che invece interessa (le palle da erba) o potrebbe interessare (racchette, asciugamani e polo) a noi.

Normalmente se andiamo in un negozio di articoli sportivi troviamo all'ingresso le indicazioni (Tennis - 1° piano, Basket - 2° piano - Arti Marziali - 3° piano) il che ci aiuta a non perdere tempo e a trovare immediatamente l'oggetto (o gli oggetti) del desiderio.

Ci scuserete per la banalità dell'esempio ma vi pregheremmo di tenerlo presente nella prosecuzione di questo capitolo in quanto riteniamo sia abbastanza esplicito nell'introdurre il concetto del data warehouse e del perché la vostra azienda ne ha probabilmente bisogno.

Quando un'azienda si trova nella situazione 1 (magazzino senza commesso) probabilmente ha bisogno di un data warehouse, se si trova nella situazione 2 è altrettanto probabile che ne abbia già uno.

Nel magazzino (il sistema informativo aziendale) ci sono tutte le informazioni che servono (TUTTI gli attrezzi di TUTTE le marche per TUTTI gli sport) per prendere le decisioni, noi siamo interessati ad uno specifico subset informativo (siamo appassionati giocatori di tennis) e siamo alla ricerca di una o più informazioni specifiche (le palle da erba) ma non siamo in grado di trovarcele da soli (il commesso non c'è o non ha voglia di aiutarci) a meno di perdere tantissimo tempo e senza alcuna certezza di raggiungere il nostro obiettivo (il magazzino è organizzato in base a criteri a noi totalmente ignoti e comunque incomprensibili).

Ma quali sono gli oggetti di cui noi andiamo alla ricerca nel magazzino rappresentato dal sistema informativo aziendale?

1.1.1. Sappiamo … ? (Domande imbarazzanti!)

- Quali sono le 5 filiali più *performanti* per regione?
- Quali sono i 10 clienti più redditizi per ogni filiale negli ultimi 3 mesi?
- Qual é il margine lordo per famiglia di prodotto e fascia di clientela delle ultime 12 settimane?
- Qual è la performance del patrimonio gestito per area manager e per cliente?

Probabilmente no, ma perché dovremmo avere queste informazioni?

- i concorrenti potrebbero avere un livello di conoscenza della clientela migliore del nostro e di conseguenza avrebbero la possibilità di fornire prodotti e servizi personalizzati;
- i concorrenti avrebbero migliori informazioni nel guidare il cambiamento o reagire alla dinamica del mercato;
- diventeremmo meno efficienti nel fornire i prodotti ed i servizi sul mercato;
- diventeremmo meno efficaci dei nostri concorrenti nel gestire i rischi insiti nel business

Come in più di un occasione ha ricordato il Presidente degli Stati Uniti Bill Clinton "l'Informazione è Potere", il che, traslato in un contesto di business significa che l'informazione si trasforma in un vantaggio competitivo. L'aspetto più bizzarro di tutto ciò è che il tipo di informazione di cui sopra è disponibile in azienda, ma non viene sfruttata, perché?

1.1.2. Perché costruire un data warehouse?

I motivi per cui un'azienda decide di dotarsi di un'architettura di data warehousing sono di vari tipi. Vediamone alcuni.

Ricerca di opportunità di crescita. In generale possiamo pensare al data warehouse come ad un sistema di supporto

decisionale avanzato. Ne segue che come tale deve guidare i processi di cambiamento in atto nell'azienda, inclusa la crescita della stessa. Questo può comportare la ricerca di opportunità per migliorare la penetrazione di mercato (nuove nicchie,cross-selling, etc.) oppure il miglioramento del servizio al cliente per creare *opinion leader* o semplicemente arricchire l'offerta.

Il data warehouse, con la sua capacità di effettuare analisi di dettaglio su grandi masse di dati rappresenta uno strumento unico nel suo genere per conoscere a fondo i nostri clienti ed il loro comportamento di acquisto.

Creazione di efficienza interna. Uno dei problemi più sentiti e diffusi nelle aziende dal punto di vista decisionale è l'affidabilità e la certificazione dei dati utilizzati per prendere le decisioni. Abbiamo visto in molte aziende situazioni kafkiane in cui nelle riunioni ognuno porta con sé la 'sua' versione della verità, verità che quasi mai coincide o anche solo assomiglia alla verità degli altri. Ciò avviene perché non esiste un'unica versione delle informazioni utilizzate per prendere le decisioni. Ognuno estrae le informazioni dai sistemi informativi disponibili in maniera diversa dagli altri usando un proprio criterio di estrazione, un proprio criterio di elaborazione ed un proprio criterio di interpretazione dei dati.

Un altro aspetto importante è l'interrelazione tra le informazioni utilizzate per prendere le decisioni Il data warehouse documenta e spiega tali interrelazioni favorendo il lavoro in team, ove una situazione destrutturata (informazioni estratte dai sistemi dipartimentali in maniera slegata) genera invece conflitti e visioni difformi della realtà.

Un effetto virtuoso di tali interrelazioni e del teaming che favoriscono è la capacità dell'organizzazione di muoversi in maniera coordinata in risposta a stimoli provenienti dal mercato.

Contenimento dei costi. Questo tipo di effetto si ha solo quando il sistema di data warehousing è a regime e comunque dipende dal tipo di focalizzazione che il sistema ha. Intanto va chiarito che non è assolutamente vero che il data warehouse è strumento del controllo di gestione: il controllo di gestione può essere un'area applicativa del data warehouse o esserne uno degli utenti, è scorretto pensare al data warehouse come unica soluzione possibile per esigenze di reporting o di controllo dei costi. Il costo da sostenersi per implementare questa soluzione è in taluni casi spropositato rispetto alla complessità delle analisi da effettuarsi.

É comunque indiscutibile che un sistema di analisi della redditività basato su un'architettura di warehousing consente in maniera esplicita ed oggettiva di comprendere quale sia la performance dei prodotti/servizi forniti dall'azienda da un punto di vista reddituale e, nell'ambito di tale analisi, di comprendere quale sia la struttura di costo generata da tali prodotti/servizi.

Supporto alle decisioni strategiche. Un'architettura di data warehousing è nello stesso tempo un potente strumento di analisi e di sintesi. Di analisi per la capacità (se supportata da un'architettura tecnologica ben costruita e dimensionata) di cogliere informazioni significative nella massa di dati contenuta nel sistema informativo aziendale, di sintesi per la possibilità di effettuare ogni tipo di sintesi decisionale partendo proprio dai dati analitici.

Il data warehouse è nello stesso tempo uno strumento in grado di evidenziare informazioni sintetiche per i decisori aziendali di livello alto e una piattaforma di ricerca di informazioni per i cosidetti *'Knowledge Workers'*, cioè gli analisti di business che usano il data warehouse per supportare il proprio lavoro analitico e decisionale.

Ne segue che gli utilizzi a scopo strategico e decisionale del data warehouse vanno dal posizionamento globale dell'azienda al posizionamento dei prodotti/servizi nel business di riferimento, dall'analisi del profilo della clientela all'analisi di redditività dei business in cui l'azienda opera.

1.1.3. Le barriere all'uso efficace dell'informazione

Come già accennato precedentemente, in azienda tutti i dati utili per prendere decisioni sono disponibili. Il problema consiste nel fatto che tali dati si trovano all'interno dei sistemi transazionali, il che ne rende impossibile lo sfruttamento a fini analitici.

Ma perché i dati transazionali non sono utilizzabili per prendere le decisioni? Perché esiste la necessità di creare un data warehouse, che prima di ogni altra cosa è una (secondo molti) pericolosa ridondanza di dati all'interno del sistema informativo aziendale?

Primo. Non è detto che il sistema informativo aziendale si basi su un solo sistema operativo, su un solo DBMS, su un solo server e sia in un solo posto fisico. Questo è vero soprattutto nel settore finanziario, i cui sistemi informativi hanno una serie di caratteristiche peculiari rispetto a quelli di altri settori. Le banche sono state fra le prime aziende a sfruttare in maniera massiccia l'informatica e quindi il processo di stratificazione dei sistemi informativi bancari ha avuto ad oggi una durata almeno tripla rispetto a quella di tutti gli altri settori. Tale stratificazione ha generato una babele di piattaforme, sistemi e linguaggi non sempre e non necessariamente fra loro integrati. A peggiorare la situazione si aggiunge il fatto che i sistemi informativi bancari, più di quelli di altri settori, sono di tipo *custom* (cioè sviluppati 'in house' a fronte di esigenze ritenute assolutamente specifiche), il che ovviamente non ne facilita l'integrazione.

Secondo. Un utente normale, senza particolari competenze tecniche difficilmente è in grado di districarsi all'interno di sistemi informativi complessi, basati su linguaggi astrusi, soprattutto se tale sistema è documentato in maniera frammentaria e comunque non orientata all'utenza di business.

Terzo. Molto spesso le analisi richiedono di combinare informazioni provenienti da contesti (cioè settori aziendali e quindi potenzialmente sistemi informativi dipartimentali) diversi, il che si combina drammaticamente con le difficoltà di cui ai punti precedenti.

Quarto. I dati nei sistemi informativi transazionali non sono organizzati per effettuare analisi di business ma per supportare le transazioni.

SISTEMI TRANSAZIONALI (p.es.: ERP)	**DECISION SUPPORT SYSTEM (p.es.: data warehouse)**
Catturare i dati	Generare informazioni
Grandi volumi di transazioni semplici	Numero ridotto di query complesse
Applicazioni statiche	Applicazioni dinamiche
Automatizzano la routine	Stimolano la creatività
Obiettivo primario: **SUPPORTARE IL BUSINESS**	Obiettivo primario: **ANALIZZARE IL BUSINESS**

Tabella 1 - Sistemi transazionali vs. DSS

Credo che questa frase (abusata oltre ogni decenza da tutti coloro che a vario titolo si occupano di data warehousing) vada in questa sede spiegata un po' meglio, con qualche esempio che ne aiuti a comprendere la portata.

Solitamente un sistema informativo per la gestione dei conti correnti è focalizzato soprattutto su un'entità e su due grandezze fondamentali: il conto corrente, il saldo contabile ed il saldo liquido. Ne segue che tutte le informazioni contenute in tale sistema informativo hanno come *focus* principale il **prodotto**, e non è improbabile che un **cliente** in possesso di più di un conto corrente venga censito più volte. Questo perché da un punto di vista contabile la cosa più importante è il conto corrente. Da un punto di vista marketing invece esiste una sempre maggior 'fame' di informazioni relative al cliente ed ai suoi rapporti con la banca.

Esempio 1. Qual é il saldo medio dei clienti della provincia di Brescia? E della Lombardia? Per avere questa informazione dovremmo, tanto per cominciare sapere che Borno, Capodiponte, Piancamuno ed Edolo fanno parte della provincia di Brescia e non è detto che questa informazione sia disponibile (in modo automatico). Nel portare i dati nel data warehouse potremmo aggiungere l'informazione 'BS' e magari 'Lombardia' e questo ci consentirebbe presto e bene di effettuare tale analisi.

Esempio 2. Qual é la distribuzione per fascia di età della mia clientela? Di solito in anagrafica è disponibile solo la data di nascita del cliente. Nel data warehouse potremmo settimanalmente eseguire un programma che, scorrendo tutta l'anagrafica clienti aggiorna un ipotetico campo 'età' in maniera tale da consentirci l'analisi di cui sopra con una serie di semplici SELECT COUNT.
Esempio 3. Qual é il saldo totale dei conti correnti del signor X? Mettere in correlazione tutte le posizioni contabili del signor X è un'impresa improba in quanto non è improbabile che ogni volta che il signor X ha per esempio aperto un conto corrente sia stato censito come nuovo correntista. Nel data warehouse potremmo effettuare un'attività di *cleansing* che generi un'anagrafica clienti con un solo record per ogni persona fisica utilizzando per esempio il codice fiscale come chiave di identificazione e legando tale informazione a tutte le posizioni contabili aperte. A questo punto l'analisi di cui sopra sarebbe effettuabile con una semplice SELECT JOIN.

1.1.4. Quali benefici il data warehouse può portare al Business?

Volendo sintetizzare quanto esposto nei due precedenti paragrafi potremmo indicare i seguenti punti:

- **Miglioramento della redditività**
  - Riduzione dei costi operativi: i processi di analisi e decisione sono estremamente costosi e complessi; il data warehouse è una tecnologia facilitante tali processi e come tale ne riduce i costi; riduce inoltre il margine di errore sulle decisioni in quanto eleva il livello di informazione e quindi di consapevolezza con cui esse vengono prese;
  - Riduzione dei costi di marketing un utilizzo attento in chiave analitica del data warehouse consente di evitare costose e inutili indagini esterne e soprattutto di guidare meglio le campagne: quante iniziative di marketing si rivelano un fiasco perché intempestive o male indirizzate? E molto spesso ciò è dovuto a

mancanza d'informazione, ciò che il data warehouse primariamente fornisce;

- **Crescita della quota di mercato**
  - Sviluppo di nuovi prodotti/servizi: il data warehouse consente analisi di dettaglio sulle performance di business delle aziende e, soprattutto, di comprendere meglio i clienti e loro esigenze; ne segue che il data warehouse, utilizzato come strumento di analisi di dettaglio del mercato, guida la definizione di nuovi e migliori modi di servirlo;
  - Identificazione, segmentazione e fidelizzazione del cliente: un altro importante aspetto che può portare a miglioramenti nella quota di mercato è la possibilità di effettuare analisi di dettaglio sui clienti; dove questo concetto è particolarmente esasperato è possibile arrivare al sogno proibito di ogni direttore marketing: il marketing one-to-one.

Perché quindi il data warehouse? Perché l'informazione è l'elemento competitivo più importante e talvolta esiziale per la sopravvivenza, soprattutto in mercati dinamici e poco stabili come quelli finanziari: il data warehouse è una risposta strutturata e quindi efficace ed efficiente a tale esigenza.

1.2. Cos'è il data warehouse ?

In questo paragrafo spiegheremo cosa è il data warehouse, con la speranza che molti dei proclami del precedente paragrafo trovino una giustificazione concreta e razionale.

Il modo migliore per spiegare il concetto di data warehouse è dichiarare innanzitutto cosa NON è. Il problema principale del data warehouse è che viene considerato un concetto non univoco e ciò permette che qualunque cosa possa essere spacciata per data warehouse.

1.2.1. Cosa il data warehouse NON E'

Diffidate di e non comprate un'auto usata da chiunque vi dica che il data warehouse è una delle seguenti cose:

- **Un prodotto che può essere acquistato e installato**: questa è una credenza molto diffusa che spesso si sublima nella convinzione che comprando un 'prodotto di warehousing' si sia 'comprato il data warehouse'. La verità vera è che il data warehouse è un sistema informativo che va sviluppato da zero utilizzando, è vero, tecnologia software (e talvolta hardware) specifica, ma una volta acquistato il prodotto si è fatto si e no il 5% del lavoro, senza contare che molto spesso la scelta si rivela sbagliata o sottodimensionata perché non è stata fatta a fronte di un'attenta, seria e competente analisi delle esigenze di business.

- **Una soluzione di business in sé e per sé**: il data warehouse non va progettato e creato in quanto tale. Un'azienda non affronta un progetto di 'data warehousing' in quanto tale, ma dovrebbe decidere in maniera seria e ragionata di risolvere una specifica esigenza di business (p.es.: il Customer Care) utilizzando la tecnologia del data warehouse. Se un progetto di data warehousing parte con presupposti diversi da questi è destinato a fallire.

- **Una 'pezza' ai problemi dei sistemi transazionali**: il data warehouse è un'ottima risposta a esigenze di analisi e supporto decisionale, funzionalità supportate poco e male dai sistemi transazionali. Punto. Non deve essere utilizzato per risolvere altri problemi che non è deputato a risolvere. Alcuni esempi:
  - calcolo dei costi di produzione
  - stampa delle etichette per il mailing
  - ribaltamenti e 'spalmature' della contabilità analitica
  - stampa del conto economico aziendale (civilistico)
  - contact management
  - data entry manuale di informazioni

Se qualcuno vi dice che intende usare il data warehouse per una o più di queste cose fermatelo prima che si faccia del male: spenderà

tantissimi soldi e otterrà risultati scarsi. In realtà molto spesso chi crea confusione su cosa sia e come si fa un data warehouse sono i software vendor e le società di consulenza, gli uni nel tentativo di legare i propri prodotti a slogan e concetti che nel mercato 'tirano', le seconde nel costante tentativo di inventare qualcosa di nuovo da proporre ai clienti. Ecco che così un *report writer* qualsiasi diventa 'un prodotto OLAP' ed un database su PC usato per stampare i conti economici per prodotto diventa 'il data warehouse dell'analisi della redditività'.

Costruire un data warehouse è un'impresa lunga e complessa dove nulla va banalizzato e dove le competenze di tipo tecnologico contano esattamente quanto quelle funzionali così come quelle funzionali contano esattamente quanto quelle tecnologiche.

1.2.2. Cosa il data warehouse E'

Stabilito che il data warehouse non è nessuna delle cose illustrate sopra, proviamo a dare qualche definizione spicciola che dettaglieremo meglio nelle prossime pagine. Il data warehouse è:

- **Un processo**: prima di ogni altra cosa il data warehouse è un processo che parte dall'estrazione dei dati di base dai sistemi informativi transazionali per arrivare alla presentazione di tali dati trasformati in informazioni all'utilizzatore finale.

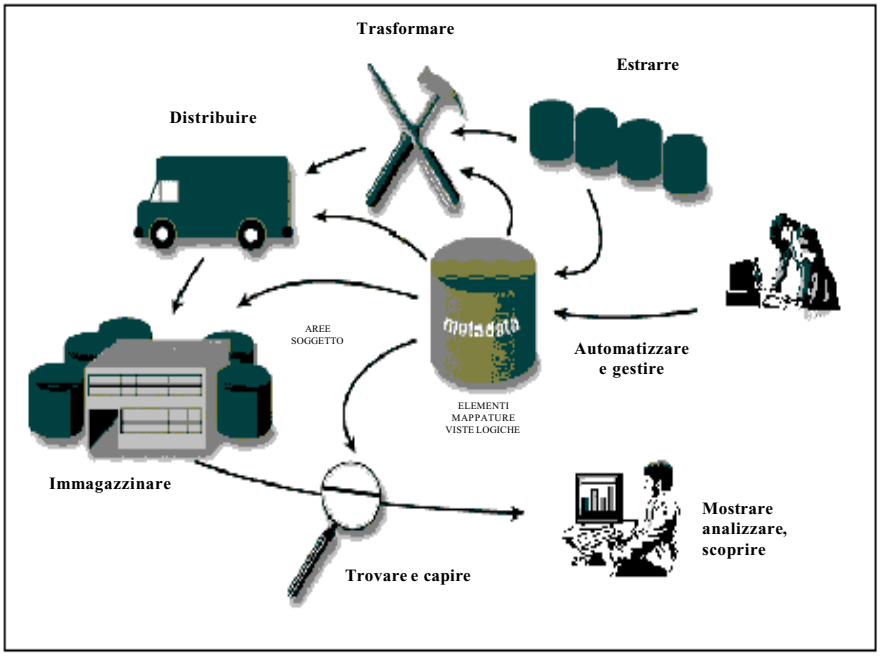


Figura 1 - Il processo del data warehouse

I principali step del processo del data warehouse sono: l'**estrazione** dei dati dai sistemi alimentanti, la **trasformazione** (integrazione, applicazione di 'business rule' e validazione) dei medesimi, la **sommarizzazione** dei dati di dettaglio per ottenere sintesi decisionali e la **presentazione** dei medesimi agli utenti. Quindi, essendo il data warehouse un processo, si può facilmente intuire che problemi o malfunzionamenti in una qualsiasi delle attività che lo compongono compromettono inevitabilmente il risultato finale.

E' perciò molto importante tenere a mente che tutte le componenti del data warehouse sono ugualmente importanti e degne di considerazione: un bel front-end che presenta dati inconsistenti e frammentari o una base dati perfettamente disegnata e alimentata ma con un front-end inaccessibile e oscuro agli utenti sono due esempi di sistemi di data warehousing non perfettamente riusciti.

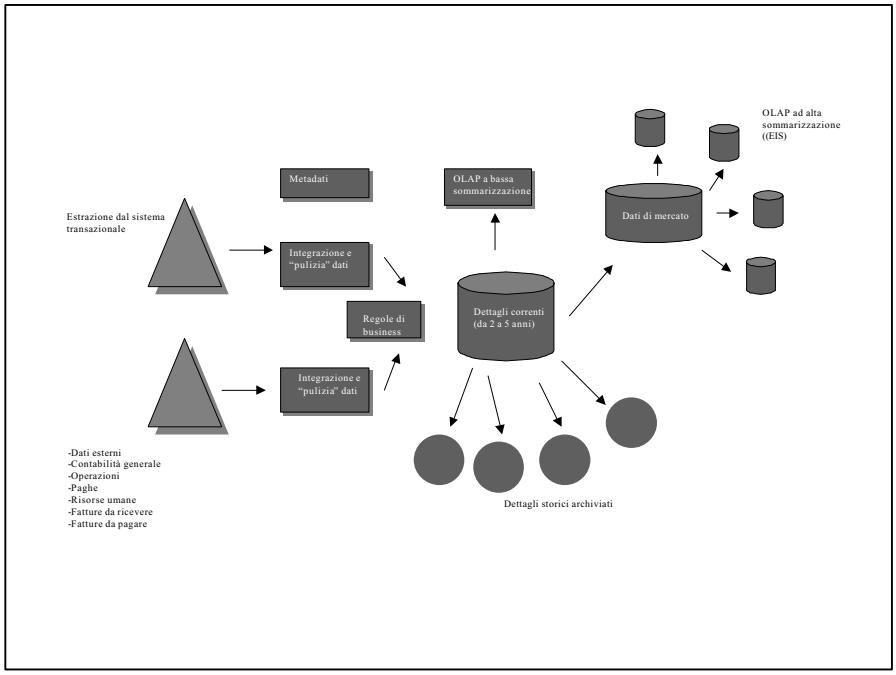


Figura 2 - Flusso e dettaglio dati nel data warehouse

- **Un'architettura**: l'aspetto architetturale è estremamente importante: il data warehouse deve essere dotato della capacità di eseguire operazioni complesse e 'pesanti' dal punto di vista dell'impegno di capacità di elaborazione; è quindi richiesta una significativa capacità di progettazione e pianificazione dell'architettura di supporto del sistema.
  Altro aspetto da non trascurare è il fatto che realizzare un data warehouse é un complesso problema di *system integration*, il che implica che i problemi architetturali sono un elemento fondamentale della costruzione di un data warehouse. Le insidie tecnologiche di un progetto di warehousing sono nascoste ovunque. Alcuni esempi:
  - il dimensionamento del o dei server fare *capacity planning* per un sistema di data warehousing è cosa estremamente complessa in quanto non è prevedibile con precisione quante risorse hardware saranno necessarie. L'unica via per risolvere il problema è puntare su piattaforme hardware scalabili e 'pompare' memoria e

processori con il crescere delle esigenze e del numero di utenti;

- le memorie di massa: anche qui esiste un problema simile a quello evidenziato al punto precedente. L'utilizzo di server OLAP (database server multidimensionali, 'generatori di cubi' per intenderci) genera molto spesso un'esplosione imprevedibile di spazio su disco; anche in questi casi l'approccio migliore consiste nel dimensionare progressivamente l'hardware a fronte del crescere delle esigenze;
- le reti di comunicazione: in funzione dei prodotti software che vengono scelti vanno considerate le implicazioni sui sistemi di comunicazione. L'utilizzo di applicazioni *thin client* (cioè con la maggior parte delle elaborazioni che avvengono dal lato server) comportano un basso livello di utilizzo delle reti, all'opposto applicazioni *fat client* (cioè con la maggior parte delle elaborazioni che avvengono dal lato client) comportano un aggravio significativo del traffico di rete.
  Altro elemento da considerare è il traffico esistente sulla rete di comunicazione. Questo aspetto, dovuto all'esistenza di altre applicazioni, deve essere attentamente valutato in fase di analisi;
- la compatibilità tra gli elementi dell'architettura: è questo l'aspetto più problematico anche perché l'unico veramente imprevedibile e ingestibile a priori. Può ad esempio capitare che esistano incompatibilità tra il database server e lo strumento di front-end o addirittura tra l'hardware di base ed alcune componenti applicative.

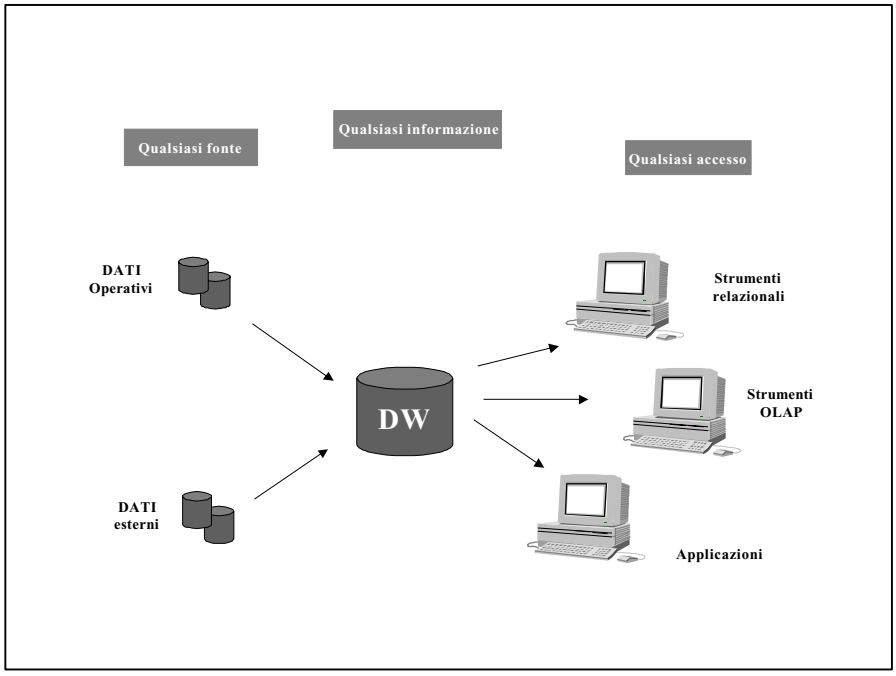


Figura 3 - L'architettura logica del data warehouse

- **Un motore di calcolo analitico**: tutti i processi di calcolo che avvengono all'interno dell'architettura del data warehouse supportano le attività di analisi e decisione. Questo significa che il data warehouse NON genera nuovi dati ma consente di analizzare meglio quelli provenienti dai sistemi transazionali trasformandoli in informazioni.

- **Una base per i processi decisionali aziendali**: un data warehouse che non viene utilizzato per prendere le decisioni è destinato a morire. Questo implica che il primo criterio che deve guidare la realizzazione di un data warehouse deve essere la sua capacità di analisi e sintesi dei dati provenienti dai sistemi alimentanti.

1.2.3. L'architettura del data warehouse

Come anticipato il data warehouse è, prima di ogni altra cosa, un processo organizzativo fortemente basato sulla tecnologia il cui obiettivo è supportare un altro importante processo aziendale: il processo decisionale. Per svolgere questo suo compito il data warehouse viene strutturato su quattro livelli architetturali:

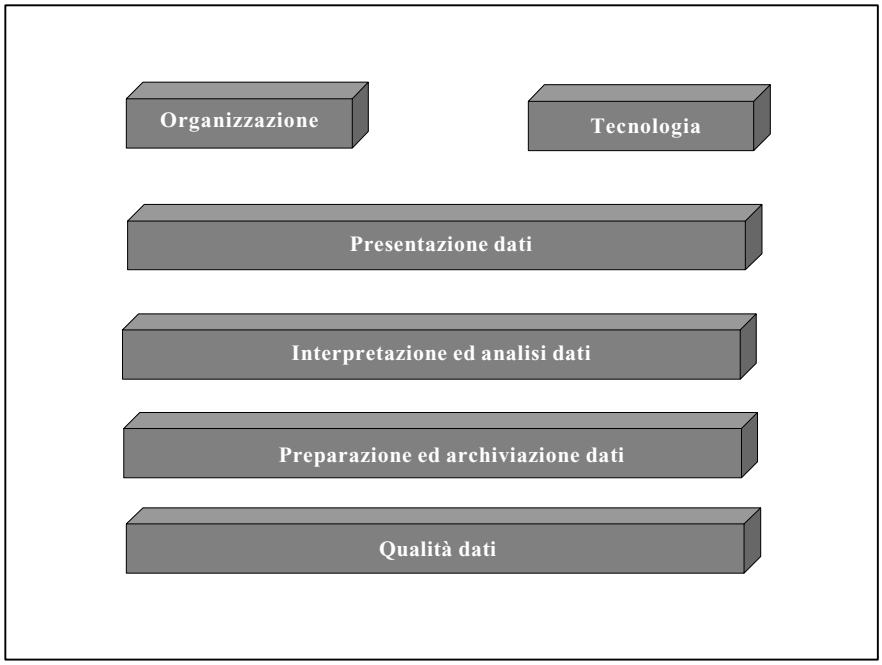


Figura 4 - Il data warehouse: organizzazione e tecnologia

- **Qualità dei dati**: è il livello che presiede all'acquisizione e validazione dei dati nel warehouse;
- **Preparazione e 'stoccaggio' dati**: è il livello che presiede alla delivery dei dati agli utenti e alle applicazioni analitiche;
- **Interpretazione e analisi dati**: è il livello, ad elevato valore aggiunto, che presiede alla trasformazione dei dati in informazioni ad elevato valore strategico;
- **Presentazione dati**: è il livello, a basso valore aggiunto, che presiede alla presentazione finale agli utenti delle informazioni e quindi delle risposte cercate.

Questi quattro livelli 'operativi' del data warehouse possono esistere sotto due condizioni fondamentali:

- l'esistenza di un'adeguata **organizzazione** di supporto al processo, con ruoli e responsabilità chiaramente definiti. Esattamente come e forse più delle applicazioni transazionali un sistema di *decision support* necessita di figure organizzative con la responsabilità di manutenerlo, soprattutto in chiave evolutiva, per far sì che esso sia costantemente allineato alle esigenze degli utenti di business, condizione necessaria e sufficiente perché continui ad esistere;
- il giusto rilievo alla **tecnologia** di supporto al processo, composta di scelte equilibrate e basate sulle esigenze funzionali del processo stesso. La tecnologia è particolarmente cruciale per il data warehouse, date le problematiche di *system integration* che esso porta con sé. La gestione costante della variabile tecnologica è uno dei fattori critici di successo del data warehouse, a partire dalle scelte iniziali per arrivare alla gestione operativa degli upgrade e degli ampliamenti della piattaforma.

Vediamo ora, da un punto di vista fisico, come è fatta un'architettura di data warehousing. Va premesso che da un punto di vista architetturale il data warehouse è un sistema **periferico**, non è cioè fisicamente residente sul sistema informativo centrale. Il motivo di ciò va ricercato nel peculiare tipo di attività svolto: una piattaforma di tipo transazionale è maggiormente orientata all'esecuzione costante di operazioni di update, ragione per cui l'ottimizzazione viene fatta soprattutto sull'I/O; una piattaforma di decision support invece deve essere ottimizzata per effettuare un numero limitato di query particolarmente complesse e CPU-consuming. L'unica eccezione a tale regola può essere rappresentata da soluzioni di tipo mainframe, ove la possibilità di definire macchine virtuali all'interno della stessa macchina fisica rende possibile la coesistenza sullo stesso server fisico delle applicazioni transazionali e delle applicazioni di *decision support*.

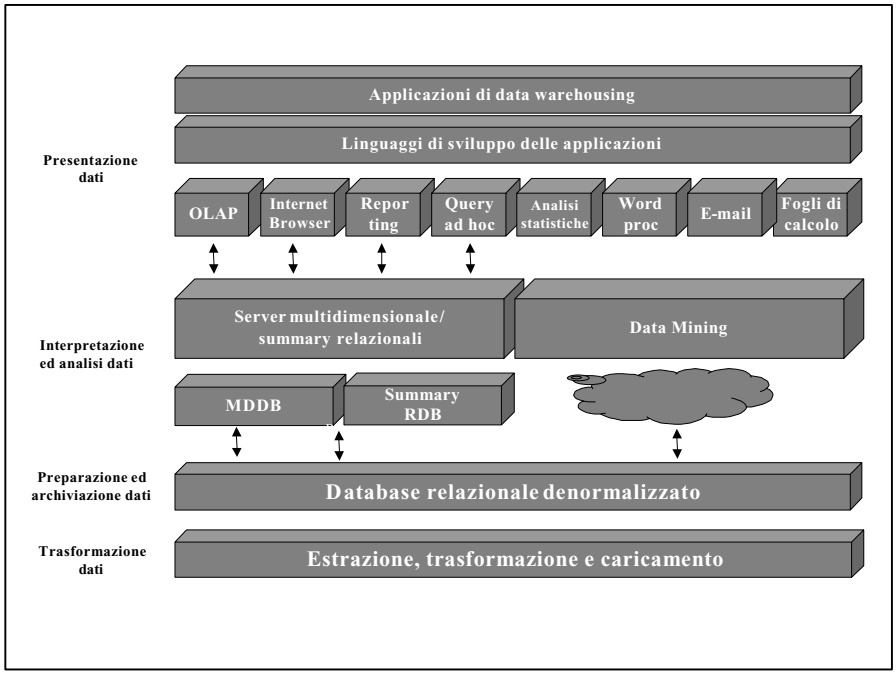


Figura 5 - Architettura fisica di un sistema di data warehousing

DATA TRANSFORMATION LAYER

L'architettura di data warehousing vera e propria inizia dallo strato denominato **data transformation**, cioè dall'insieme di applicazioni che svolgono l'attività di estrazione, trasformazione e caricamento dei dati dai sistemi transazionali alimentanti verso il data warehouse.

La fase di **estrazione** dei dati dai sistemi alimentanti viene nella maggior parte dei casi implementata utilizzando i linguaggi proprietari delle piattaforme alimentanti. Si tratta per lo più diquery ad hoc, parametrizzate per quanto riguarda l'arco temporale, eseguite periodicamente (giornalmente, settimanalmente o mensilmente) solitamente nei momenti di minor attività del sistema (la notte od il weekend). Esistono tuttavia diverse soluzioni software, basate anche su interfacce grafiche Windows-like che consentono di costruire estrazioni anche da piattaforme proprietarie mediante semplici operazioni di drag & drop.

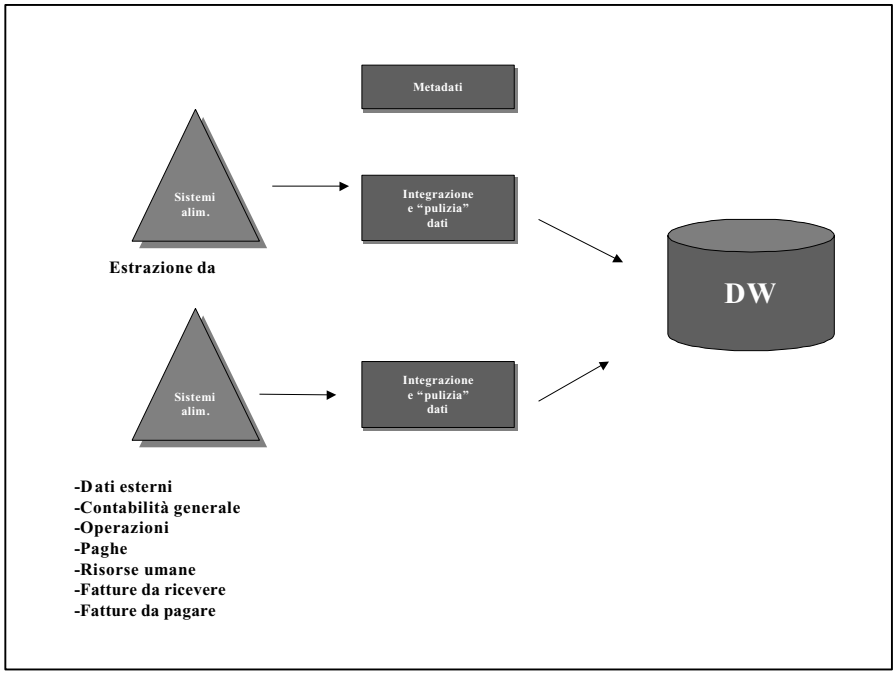


Figura 6 - Estrazione

La fase di **trasformazione**, quella a più elevato valore aggiunto fra le tre contenute in questo layer applicativo, applica regole di integrazione, trasformazione e *cleansing* (business rule) ai dati estratti dai sistemi alimentanti. E' in questolayer che molto spesso si gioca la credibilità dei dati del data warehouse presso gli utenti. Nella maggior parte dei casi i dati estratti dai sistemi transazionali sono o incompleti o comunque non adatti a prendere decisioni in quanto non coerenti con le analisi da effettuarsi (cfr. Esempi § 1.3).

In taluni casi le operazioni di trasformazione possono risultare nella casistica del *reject*: cioè dell'impossibilità, salvo intervento umano, di accettare parte del flusso alimentante per l'eccessiva e non risolvibile 'impurità' dei dati alimentanti.

Le trasformazioni possono essere di vari tipi:

- **Codifiche inconsistenti**. Lo stesso oggetto è codificato in modi diversi a seconda del sistema alimentante. In fase di trasformazione ogni flusso alimentante andrà ricodificato in

base ad un unica codifica convenzionale definita per il data warehouse;

- **Unità di misura/formati inconsistenti**. E' il caso in cui la stessa grandezza viene misurata con unità di misura o rappresentata con formati differenti a seconda del sistema alimentante di provenienza. In fase di trasformazione ogni flusso alimentante andrà convertito in un'unica unità di misura convenzionale per il data warehouse;
- **Denominazione inconsistente**. E' caso in cui a seconda del sistema transazionale alimentante lo stesso oggetto (di solito un dato) viene denominato in modo diverso. Solitamente il dato all'interno del warehouse viene identificato in base alla definizione contenuta nei metadati del sistema;
- **Dati mancanti o errati**. Nei tre casi precedenti le operazioni di trasformazione consistevano essenzialmente in attività di conversione, entro certi limiti automatizzabili. In questo caso, invece, l'operazione di trasformazione può richiedere l'intervento umano per risolvere casistiche non prevedibili e implementabili mediante le business rule.

  Un esempio per questa casistica (cfr. Fig. 2.7): supponiamo che il sistema stia acquisendo l'anagrafico clienti, entro il quale si trovano sia persone fisiche che persone giuridiche. Supponiamo anche che, oltre al codice e alla descrizione del cliente siano disponibili altri due campi: un campo booleano 'persona giuridica' (valori possibili V ed F) ed un campo 'sesso' (M, F e N). Nel caso in cui arrivi un record relativo ad un cliente persona giuridica il cui campo 'sesso' sia M o 'null' è facile prevedere un automatismo di trasformazione che valorizzi il campo 'sesso' a 'N'. Nel caso in cui invece arrivi un record relativo ad un cliente persona fisica il cui campo 'sesso' sia 'N' o 'null' è impossibile per un algoritmo automatico effettuare l'operazione di cleansing; ne segue che il record verrà rigettato e dovrà essere successivamente corretto (probabilmente a mano) e rialimentato.

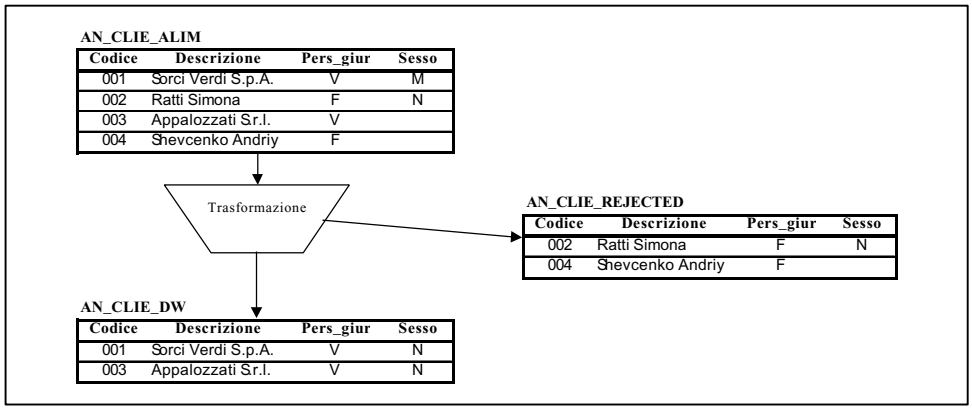


Figura 7 - Dati mancanti

DATA PREPARATION AND STORAGE LAYER

Questo livello coincide con il massimo dettaglio disponibile (in termini di dati) all'interno del sistema di data warehousing.

Una volta che i dati hanno superato il *quality layer* vengono 'stoccati' in questo livello architetturale per garantire due tipi di utilizzi:

- la creazione di sintesi informative per gli utenti (datamart e aggregazioni) mediante procedure ad hoc che vengono solitamente innescate (in termini di *update*) dalla completa esecuzione delle operazioni di estrazione, trasformazione e caricamento;
- l'esecuzione di analisi avanzate basate prevalentemente su algoritmi di tipo statistico che richiedono l'operatività sul massimo dettaglio disponibile dei dati per restituire risultati significativi.

DATA INTERPRETATION AND ANALYSIS LAYER

A questo livello dell'architettura del sistema di data warehousing troviamo oggetti tra loro molto diversi per funzione e tecnologia.

Le tre funzionalità base espletate da questo livello architetturale sono: aggregazione, analisi e interpretazione.

Aggregazione

La funzionalità di **aggregazione** provvede a costruire sintesi decisionali partendo dai dati di dettaglio presenti nel layer precedentemente descritto. Qui va fatta una importante precisazione architetturale.

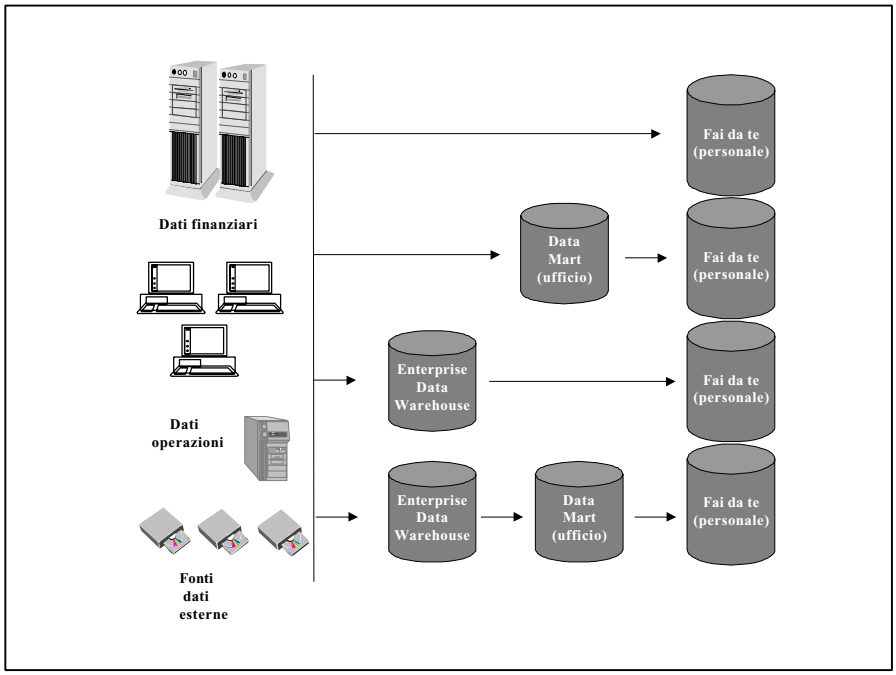


Figura 8 - Architetture di data warehousing

In una situazione in cui non esiste il data warehouse (figura 8 - caso 1) gli utenti sono costretti ad accedere ai sistemi legacy per ottenere le informazioni loro necessarie. Abbiamo già descritto nel primo paragrafo le implicazioni di tale situazione.

In taluni casi (figura 8 - caso 2) si può decidere di estrarre dai sistemi legacy una o più sintesi (data mart) per gli utenti che effettueranno l'analisi su di esse. In questa situazione, anche se la tecnologia e l'architettura assomigliano molto a quelle di un data warehouse, l'impossibilità di arrivare a dati di dettaglio superiore di quello delle sintesi disponibili (drill-through) ne riduce la potenza informativa.

Peraltro il data warehouse non va necessariamente considerato come una base dati a cui tutti gli utenti accedono liberamente per le proprie analisi (cfr. Figura 8 - caso 3). Questo può essere vero ove gli utenti siano particolarmente addestrati e, comunque sia, ha delle serie controindicazioni in quanto le risorse hardware necessarie per supportare un elevato numero di utenti che eseguono *query multijoin* sono difficilmente prevedibili e pianificabili. Molti presunti progetti di Warehousing falliscono proprio perché ci si limita a 'portare dentro i dati' senza però di fatto renderli disponibili agli utenti meno esperti.

La situazione ideale (figura 8 - caso 4) è quella in cui esiste un data warehouse centrale che contiene tutte i dati al minimo livello di dettaglio richiesto per effettuare analisi avanzate (p.es.: data mining) e per costruire aggregazioni per tutti gli utenti. In questo caso i data mart possono essere o **tematici** (cioè contenenti tutte le informazioni riguardo un certo soggetto - le vendite per esempio) o per **gruppi specifici di utenti** (data mart per famiglie di prodotti, ognuno dei quali generato per i product manager responsabili).

Questa strategia architetturale fa del data warehouse un vero processo di *information delivery*, ove la richiesta di altre e diverse sintesi decisionali comporta non già la costruzione di altri flussi di alimentazione ma piuttosto la creazione di altri data mart. Come avrete compreso lo sviluppo di nuovi flussi generanti nuovi data mart è un'attività routinaria di gestione del data warehouse. La differenza con quanto si dovrebbe fare utilizzando i sistemilegacy (figura 8 - caso 1) è essenzialmente di costo: generare un nuovo data mart all'interno di un'architettura di warehousing ha costi e tempi di sviluppo e di controllo qualità dei dati nettamente inferiori.

Analisi e interpretazione

La funzionalità di **analisi** consente di effettuare indagini sulle aggregazioni costruite dal sistema. Tipicamente le funzionalità di analisi di un data warehouse sono supportate da tecnologia di tipo **OLAP** (On-Line Analytical Processing).

L'OLAP è essenzialmente un approccio tecnologico ai processi decisionali che si focalizza sull'analisi dimensionale delle informazioni. Le sue caratteristiche principali sono:

- **é orientato agli utenti di business**: il business è fatto a *dimensioni* non a *tabelle* e chi analizza e tenta di comprenderlo ragiona appunto per dimensioni (prodotto, canale, cliente, etc.); è per questo che una volta intuiti i due concetti fondamentali (**dimensione** e **gerarchia**) qualsiasi utente di business è in grado di utilizzare uno strumento OLAP;
- **é pensato per la risoluzione di problemi non strutturati**: a differenza dei tradizionali strumenti di reporting che presentano già le risposte preconfezionate, gli strumenti OLAP stimolano le domande e consentono analisi dicausa-effetto. Ciò avviene grazie alla loro struttura che permette la navigazione tra le informazioni utilizzando le gerarchie e le relazioni tra le informazioni stesse come 'sentieri';
- **si focalizza sulle informazioni** i motori OLAP non sono di per sé strumenti di presentazione delle informazioni ma architetture ottimizzate di data storage e navigazione; ne segue che tutto ciò che un utente trova in questo ambiente sono solo le informazioni di cui ha bisogno, organizzate secondo la logica delle dimensioni di analisi di business;
- (conseguentemente) **crea efficienza**: ovviamente il risultato netto di tutto ciò è l'efficienza creata da questi sistemi con la loro capacità di andare dal generale al particolare e di aiutare l'utente a trovare l'informazione necessaria in base a percorsi logici e non 'scartabellando'.

Le funzioni di base di uno strumento OLAP sono:

- **Slicing**: è l'operazione di rotazione delle dimensioni di analisi. È un'operazione fondamentale se si desidera analizzare totali ottenuti in base a dimensioni diverse o se si vogliono analizzare aggregazioni trasversali;

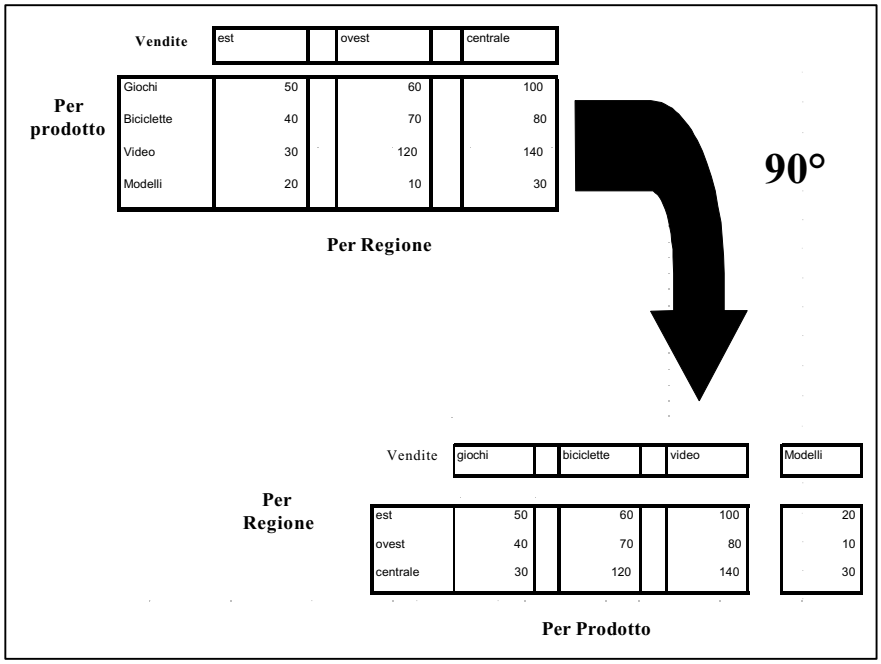


Figura 9 - OLAP: operazione di 'slicing'

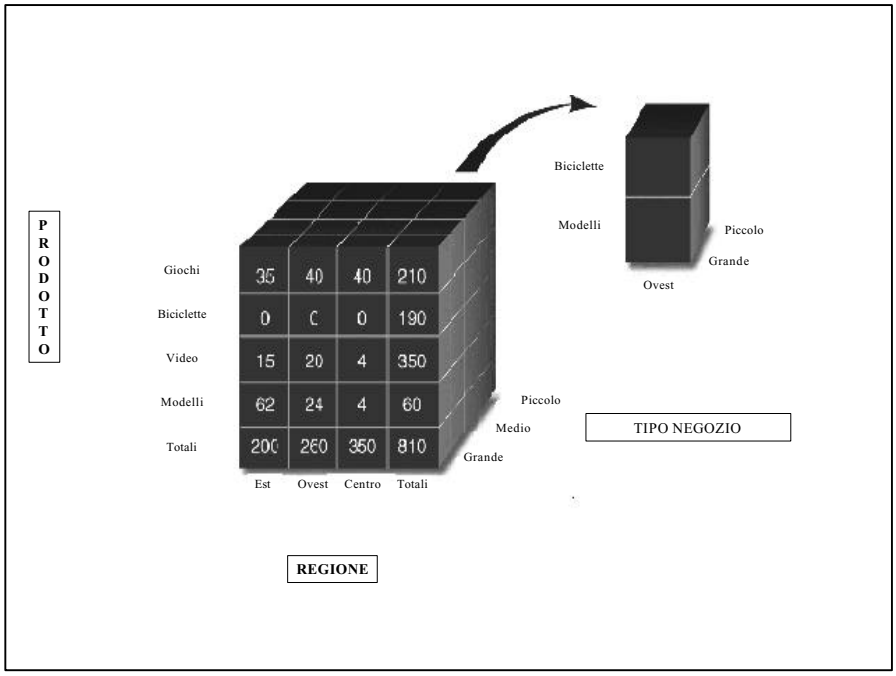


Figura 10 - OLAP: l'operazione di 'dicing'

- **Dicing**: è l'operazione di 'estrazione' di un subset di informazioni dall'aggregato che si sta analizzando.
  L'operazione di dicing viene eseguita quando l'analisi viene focalizzata su una 'fetta del cubo' di particolare interesse per l'analista. In taluni casi l'operazione di dicing può essere 'fisica' nel senso che non consiste solo nel filtrare le informazioni di interesse ma magari nell'estrarle dall'aggregato generale per distribuirne i contenuti;
- **Drill-down**: è l'operazione di 'esplosione' del dato nelle sue determinanti. L'operazione di drill-down può essere eseguita seguendo due tipologie di 'sentiero': la **gerarchia** costruita sulla dimensione di analisi (p.es.: passaggio dalla famiglia di prodotti all'insieme dei prodotti che ne fanno parte) oppure la **relazione matematica** che lega un dato calcolato alle sue determinanti (p.es.: passaggio dal margine al ricavo e costo che lo generano). Si può comprendere l'importanza di tale operazione ai fini analitici in termini di comprensione delle determinanti di un dato;
- **Drill-across**: è l'operazione mediante la quale si naviga attraverso uno stesso livello nell'ambito di una gerarchia. Come evidenziato precedentemente il passaggio dalla famiglia di prodotti alla lista dei prodotti è un'operazione di drill-down, il passaggio da una famiglia ad un'altra famiglia è un'operazione di drill-across;
- **Drill-through**: concettualmente simile al drill-down, è l'operazione mediante la quale si passa da un livello aggregato al livello di dettaglio appartenente alla base dati normalizzata. Molti software vendor proclamano che i loro prodotti hanno la capacità, mediante l'operazione di drill-through, di passare dal data warehouse ai sistemi transazionali alimentanti (p.es.: passaggio dal dato di vendita aggregato al dettaglio fattura in contabilità clienti). Tale operazione, ancorché tecnicamente fattibile sotto una serie di condizioni abbastanza rilevanti, è poco sensata per le problematiche di security e di performance indotti nei sistemi transazionali stessi.

I punti deboli degli strumenti OLAP sono:

- **Inaccessibilità/difficoltà ad accedere al livello atomico del dato**: gli strumenti OLAP funzionano molto bene su dati di sintesi, non è conveniente usarli su dati eccessivamente analitici;
- **Sistemi di backup / restore / security / rollback non molto sofisticati o inesistenti**: pur essendo in molti casi dei motori database, gli strumenti OLAP non hanno ancora raggiunto il livello di completezza dei database relazionali, principalmente perché, a differenza di questi ultimi, non hanno un paradigma concettuale di riferimento come la teoria di Codd, ma sono soggetti alle interpretazioni dei diversi software vendor;
- **Richiede una struttura denormalizzata per funzionare in maniera efficiente**: i motori OLAP generano grandi masse di dati per il semplice fatto che per migliorare le performance di accesso sono costretti a memorizzare chiavi ridondanti e sommarizzazioni;
- **(Potenziale) significativa generazione e proliferazione di codice SQL**: nel caso in cui il database su cui vengono effettuate le analisi OLAP non sia multidimensionale (MOLAP) ma sia relazionale (ROLAP), ognuna delle operazioni sopra descritte (slicing, dicing, drilling) comporta la generazione e l'esecuzione di query SQL estremamente complesse e richiedenti grandi capacità di elaborazione.

Una considerazione a parte meritano gli strumenti di analisi avanzata e interpretazione disponibili per il data warehouse: per tale argomento si rimanda ad altro capitolo del presente testo.

DATA PRESENTATION LAYER (DW APPLICATIONS)

Questo livello, ingiustamente considerato il più importante da chi pensa che costruire un sistema di decision support voglia dire disegnare degli spettacolari report layout, contiene tutti i sistemi di presentazione delle informazioni agli utenti.

I sistemi appartenenti a questo layer architetturale possono essere raggruppati in tre grandi categorie:

- **strumenti specialistici di Business Intelligence**: in questa categoria, molto vasta in termini di soluzioni presenti sul mercato, troviamo strumenti di query building, strumenti di navigazione OLAP (OLAP viewer) e, in un'accezione molto lata del concetto, anche i Web browser, che per la verità stanno diventando sempre più l'interfaccia universale per diverse applicazioni;
- **strumenti di Office Automation**: sempre più i software vendor presenti con le loro soluzioni nellayer architetturale precedente indicano come soluzioni difront-end gli ordinari strumenti del lavoro quotidiano, come word processor e fogli elettronici. E' ovvio che per molti utenti poco esperti che si avvicinano per la prima volta al data warehouse questa sia una soluzione psicologicamente rassicurante in quanto non sono costretti ad imparare nuovi e complessi strumenti. Il problema consiste nel fatto che tale soluzione è ideale per questioni di produttività ed efficienza (invece di importare una tabella da uno strumento Business Intelligence la carico direttamente nel mio word processor), lo è meno per l'utilizzo intensivo del data warehouse, dal momento che questi strumenti, in tale caso hanno, significativi limiti architetturali e funzionali;
- **strumenti di grafica e publishing**: in questo caso ciò che prevale ancora una volta è una considerazione di efficienza e produttività: gli strumenti Business Intelligence hanno la capacità di generare grafici e tabelle per i propri utenti, la soluzione in oggetto serve sostanzialmente ad evitare inefficienti doppi passaggi.

1.3. Valore strategico e valore operativo del data warehouse nelle banche e nelle istituzioni finanziarie

Prima di approfondire quale sia il valore del data warehouse emerge la necessità di analizzare quali siano le forze e le cause che hanno portato le istituzioni finanziarie a dare maggiore rilievo al

valore dell'informazione e della conoscenza. A tale riguardo risulta importante fare una breve panoramica sui cambiamenti strutturali che hanno inciso sulle istituzioni finanziarie e in particolare sulle banche e la conseguente risposta delle stesse.

1.3.1. Cambiamenti strutturali nel sistema finanziario

Lo scenario di riferimento nel quale si trovano ad operare le istituzioni finanziarie e, in particolare, l'aumentato livello di competizione che caratterizza il settore bancario, hanno spinto il sistema finanziario ad un grado sempre maggiore di efficienza, funzionalità e flessibilità operativa. In particolare, il contesto competitivo è mutato rapidamente nell'ultimo decennio sotto la spinta di fattori di cambiamento legati principalmente alle attese della clientela, ai nuovi prodotti e servizi, ai canali distributivi, nonché all'ingresso di nuovi attori sul mercato.

In particolare, i fattori che più di ogni altri hanno inciso in maniera decisa sul cambiamento strutturale possono essere così descritti:

- **Globalizzazione dei mercati** favorita dalla riduzione delle barriere all'entrata e dalla cooperazione economica internazionale. Tale fenomeno risulta più accentuato in Europa grazie all'avvento della moneta unica europea.

- **Deregulation** che ha portato alla creazione, per esempio negli Stati Uniti, di banche nazionali attraverso l'integrazione di banche focalizzate su uno specifico Stato. Tale fenomeno è in corso in Europa e specialmente in Italia dove stiamo assistendo ad una forte concentrazione bancaria.

- **Destabilizzazione della struttura di base**, che ha inciso sulla maggiore attenzione per la ricerca di nuovi servizi a valore aggiunto rispetto alle attività finanziarie tradizionali. Tale fenomeno ha comportato per le istituzioni finanziarie un

cambiamento culturale non indifferente, portando le stesse ad essere sempre meno società finanziarie e sempre più società di servizi.

- **Aumento della concorrenza**, sia fra le stesse istituzioni finanziarie al fine di acquisire una leadership su uno specifico mercato, che per le entrata di nuovi operatori economici, sempre più aggressivi grazie all'utilizzo di nuova tecnologia.

- **Aspettative della clientela**, da acquisto prodotti a soluzione di bisogni. Tale fenomeno ha portato le istituzioni finanziarie e, in special modo, le banche a prestare sempre più attenzione alla conoscenza delle esigenze della clientela e alla gestione della relazione con la stessa.

- **Evoluzione tecnologica** da un uso quantitativo ad uso qualitativo al fine di raggiungere l'obiettivo di maggiore flessibilità, migliore controllo dei costi, migliore interazione con i clienti e maggiore allineamento alle esigenze di business. La superiorità tecnologica sarà uno dei fattori principali per il successo futuro. Questo perché con la tecnologia si riuscirà a gestire in maniera efficiente ed efficace per esempio: i clienti, i prodotti, le competenze manageriali, la multi-canalità. Basti pensare alle nuove opportunità che il canale Internet può portare alle istituzioni finanziarie.

- **Governance aziendale**, guidata dalla redditività sul capitale di rischio e dalla creazione di valore per l'azionista. Tale fenomeno oltre a guidare la strategia aziendale nella ricerca di una sempre maggiore redditività, porta le istituzioni finanziarie a dotarsi di strumenti di controllo della performance aziendale, che possono coniugare l'obiettivo di semplicità e snellezza con quello di tempestività.

Tali cambiamenti strutturali hanno di fatto comportato per le aziende di questo settore la ricerca di nuove architetture finanziarie che potessero allineare i processi operativi al nuovo contesto

ambientale, attraverso una specializzazione dei ruoli, *produttivi e distributivi*, e una maggiore focalizzazione sulla clientela.

La specializzazione dei ruoli ha portato le diverse aree aziendali a focalizzarsi sul loro specifico core business e, conseguentemente, è stata abbandonata la logica del 'tutto a tutti ovunque', operando solo nelle combinazioni di prodotti/segmenti/canali di clientela in cui si ha già e/o si ritiene di poter acquisire una posizione di leadership; in una logica di riduzione di costi, miglioramento della qualità del servizio, ampliamento della gamma prodotti e efficace gestione dei rischi.

La focalizzazione sulla clientela e sui bisogni della stessa ha portato le istituzioni finanziarie a "reinventare" l'organizzazione aziendale e ad investire fortemente sull'innovazione tecnologica al fine di:

- anticipare i bisogni finanziari della clientela;
- strutturare l'azione di vendita sulla base di attività pianificate;
- ottenere e gestire informazioni per il targeting dell'azione di vendita;
- specializzare le risorse commerciali e responsabilizzarle su precisi obiettivi di vendita;
- monitorare l'azione di vendita e la qualità del servizio offerto.

1.3.2. La risposta delle banche e delle istituzioni finanziarie

Lo scenario sin qui delineato sta spingendo le istituzioni finanziarie ad una polarizzazione su tre ruoli prevalenti: **gruppi**

integrati, banche commerciali e gruppi assicurativi integrati verticalmente, con canali distributivi proprietari e raggio di azione internazionale; **produttori puri,** gestori globali con un'offerta specializzata di prodotti che collocano tramite rete multicanale non necessariamente proprietarie e rivolte a varie nicchie di mercato; **distributori puri**, operatori focalizzati sulla distribuzione per il mass market, che collocano prevalentemente prodotti di terzi.

Conseguentemente la scelta di uno dei tre ruoli ha imposto, comunque, alle istituzioni finanziarie il raggiungimento di posizioni di leadership nel mercato. Per poter acquisire tale obiettivo, imprescindibile per la stessa sopravvivenza, si è assistito ad una forte concentrazione tra i vari operatori Infatti, la mancanza di una forte massa critica e di una quota di mercato rilevante non consente all'istituzione finanziaria di poter svolgere un ruolo attivo sul mercato e quindi, sulla base anche degli elementi di cambiamento già delineati, poter sopravvivere a medio-lungo termine.

I motivi che hanno portato alla concentrazione tra gli operatori possono essere così delineati:

- saturazione del mercato;
- incremento della quota di mercato;
- minaccia per l'entrata di nuovi attori;
- riduzione dei costi;
- ampliamento dell'offerta di servizi;
- incremento dei canali distributivi;
- offerta basata su più marchi per poter attrarre più segmenti di clientela.

In tutti i casi osservati la concentrazione ha portato le istituzioni finanziarie a separare le attività distributive da quelle produttive e, nello stesso tempo, ad organizzarsi per segmento di clientela. Tale tipologia organizzativa richiede una sempre maggiore attenzione

nella gestione, analisi e distribuzione dei dati aziendali affinché, da un lato possa essere svolto un corretto controllo direzionale e, dall'altro lato, le diverse unità di business possano poter monitorare tempestivamente le performance raggiunte e quindi poter eventualmente riallineare le azioni.

Nel contempo il valore di acquisto del cliente è mutato rispetto al passato: il cliente è diventato più esigente, e conseguentemente la focalizzazione sulla clientela è divenuta più stringente.

Il cliente, infatti, richiede alle istituzioni finanziarie capacità di fornire soluzioni, disponibilità al dialogo e all'ascolto, informazioni complete, chiare ed essenziali, velocità nell'esecuzione delle transazioni, prezzi ridotti e automazione dei servizi richiesti. Conseguentemente il valore di acquisto del cliente si concentra maggiormente sulla qualità del servizio, sulla trasparenza e sull'innovazione offerta.

Tutto ciò ha indubbiamente impatto sul modo di 'fare banca'; infatti la strategia deve essere focalizzata sull'ampliamento della gamma prodotti, sulla distribuzione multicanale, sui prezzi e sulle performance attese. Le risorse devono essere dotate di competenze specifiche e con capacità relazionali, i processi allineati alle esigenze della clientela e la struttura disegnata per segmento di clientela.

In tutti i casi acquisisce valore fondamentale l'informazione. Infatti sarà importante capire il cliente, cosa ha, cosa vuole e come possiamo soddisfarlo; capire i comportamenti e come possiamo anticipare i bisogni; capire le tendenze di mercato e come possiamo allinearci.

Quindi, dalle risposte che le istituzioni finanziarie stanno dando ai cambiamenti strutturali, emerge la necessità che le informazioni aziendali siano strutturate al fine di poterle gestire, analizzare in tutte le ottiche di analisi e in ultimo presentarle. La capacità delle istituzioni finanziarie nella gestione delle informazioni e delle conoscenze rappresenterà nel breve futuro un vantaggio competitivo difficilmente colmabile e tutto questo potrà rappresentare per le stesse un patrimonio strategico aziendale.

1.3.3. Valore strategico e operativo del data warehouse

In un contesto di mercato sempre più turbolento e competitivo e di fronte alla aumentata complessità della gestione aziendale è necessario introdurre supporti alla gestione che consentano al management di cogliere le opportunità, monitorare il comportamento dell'azienda e il raggiungimento degli obiettivi nonché identificare tempestivamente le criticità e le relative cause.

I presupposti della redditività si identificano nella capacità di prevedere, interpretare, gestire a proprio vantaggio i cambiamenti ambientali e competitivi.

I tradizionali meccanismi di controllo e analisi (sui volumi, sulle filiali, etc.) devono essere integrati da sistemi di pianificazione, di budgeting, di responsabilizzazione e di monitoraggio focalizzati su dimensioni gestionali coerenti con i fattori critici di successo.

Conseguentemente l'adattamento delle istituzioni finanziarie alle mutate condizioni esterne di mercato e interne di impresa, richiede la progettazione e la messa a punto di un sistema di governo di impresa più flessibile, più reattivo in rapporto alla sequenza accelerata delle opportunità e delle minacce proposte dall'ambiente

Il governo di impresa deve essere basato su leve adeguate a supportare la focalizzazione sui fabbisogni della clientela, la responsabilizzazione diffusa su tutti i livelli aziendali, il decentramento decisionale e la focalizzazione sulle cause dei problemi e sulla loro prevenzione, con lo scopo di recuperare redditività, controllare i rischi e ridurre i costi.

A tale riguardo il modello di reporting deve essere caratterizzato da:

- facilità di utilizzo degli strumenti;
- flessibilità e versatilità degli strumenti, con conseguente possibilità di gestire i dati secondo le viste necessarie (a seconda dei fabbisogni informativi);

- possibilità di approfondire l'indagine a livelli successivi, lungo le dimensioni selezionate;
- disponibilità di formati grafici caratterizzati da immediatezza ed efficacia;
- capacità di fornire una reportistica basata sulle eccezioni;
- omogeneità della fonte dei dati.

Tale facilità nell'accesso alle informazioni agevolerà il management a capire più agevolmente i dati presenti all'interno dell'azienda, affinché diventino informazioni utili per attuare le giuste scelte che apportino miglioramenti al business e che, quindi, possano rappresentare un patrimonio strategico aziendale.

Nel contempo risulta necessario controllare i comportamenti operativi sia interni che esterni:

- centri di responsabilità;

- clienti.

Il controllo per centri di responsabilità deve fornire le informazioni necessarie per analizzare l'efficienza e la redditività dei centri dando, non solo le configurazioni di costo e di ricavo che soddisfino le suddette analisi, ma anche mettendo in moto quei meccanismi di misurazione delle attività, dei volumi e quant'altro che renda l'analisi più completa.

Il controllo dei comportamenti della clientela deve fornire le informazioni necessarie per capire cosa vendere, come venderlo e a che prezzo venderlo fornendo, non solo i dati sui volumi e sui prezzi, ma anche tutte quelle informazioni relative a cosa acquistano, perché l'acquistano e con quale canale. Soprattutto è importante che vengano delineati i comportamenti per segmento di clientela al fine di estrapolare la combinazione prodotti/segmento/canali e quindi pianificare meglio le azioni commerciali.

In ultima analisi, quindi, ogni domanda deve trovare risposta negli indicatori relativi alle diverse aree di analisi (redditività, rischi, sofferenze, ricavi, costi, servizio al cliente, etc.) e successivamente

le aree devono a loro volta essere segmentate nelle seguenti ottiche di analisi:

- gruppo e mercato;
- organizzazione territoriale;
- area geografica;
- cliente;
- prodotto;
- area di affari;
- canale di distribuzione;
- stratificazione temporale;
- processo;
- progetto;
- responsabilità.

Affinché tutto questo possa essere realizzato sarà necessario:

- omogeneizzare e sincronizzare i dati per alimentare i singoli modelli di analisi;
- centralizzare i dati provenienti da fonti diverse (mainframe, provider esterni, etc.);
- storicizzare i dati gestiti;
- rendere i dati leggibili a tutti gli utenti;
- manipolare e presentare i dati e/o le informazioni per diversi scopi di business e da parte di varie categorie di utenti finali.

Il data warehouse appare al momento la piattaforma tecnologica candidata a risolvere in maniera brillante tali esigenze.

1.4. Gli ambiti applicativi del data warehouse nelle banche e nelle istituzioni finanziarie

Come probabilmente avrete già compreso il data warehouse è un po' come il bunge-jumping: ne parlano in molti ma lo praticano in pochi. E' anche per questo che è così difficile identificare e motivare le aree applicative del data warehouse.

Va detto che peraltro nelle banche e più in generale nelle istituzioni finanziarie gli ambiti di utilizzo sono molteplici, proprio perché tutte le aree gestionali di tali organizzazioni sono caratterizzate da volumi impressionanti di dati su cui devono essere prese decisioni strategiche.

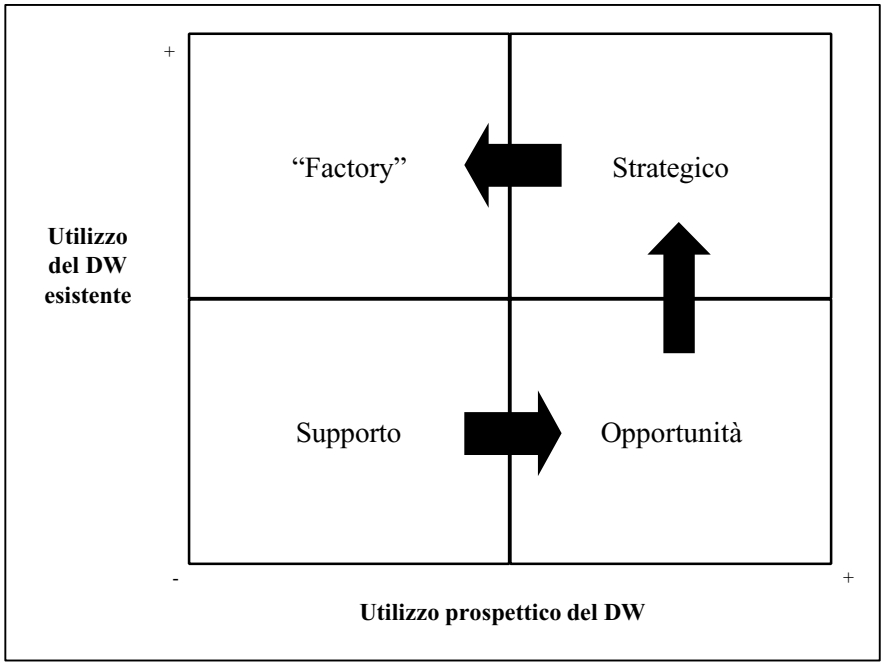
Proprio perché il data warehouse può avere un valore strategico (ed è di fatto una scelta strategica) all'interno di tali tipi di organizzazioni è fondamentale per il management definire una **strategia** per il data warehouse. La strategia per il data warehouse è essenzialmente un percorso evolutivo che porta l'azienda da applicazioni DW non ‡mission-critical' a una situazione in cui il data warehouse è una componente fondamentale del sistema informativo aziendale. Nel prossimo paragrafo comprenderemo come ciò avviene.

1.4.1. L'evoluzione nel ruolo del data warehouse

La strategia di data warehousing di un'azienda può essere classificata in base a due dimensioni fondamentali:

- utilizzo del DW esistente: livello di maturità degli utenti e delle funzioni di supporto del DW nell'utilizzo dell'esistente
- utilizzo prospettico del DW: obiettivi strategici di utilizzo del DW come piattaforma di decision support

Tutte le aziende attraversano quindi quattro fasi nella storia dell'utilizzo del data warehouse:



- la prima fase, chiamata **supporto** (basso utilizzo del DW esistente, basso utilizzo prospettico del DW), è la fase in cui si trovano le aziende che hanno fallito uno o più progetti di warehousing e non pensano, per questo o per altri motivi, di ampliarne l'utilizzo prospettico. In questa fase si possono trovare anche aziende che non hanno un DW e non stanno pensando di realizzarlo;
- la seconda fase, chiamata **opportunità** (basso utilizzo del DW esistente, alto utilizzo prospettico del DW), è la fase in cui si trovano le aziende che, pur avendo fallito uno o più progetti di warehousing o avendo semplicemente esplorato la tematica senza approfondirla, puntano a sviluppare le attività di decision support tramite il data warehouse. In questa fase possono trovarsi anche aziende che non hanno un DW ma stanno pensando seriamente di realizzarlo;
- la terza fase (alto utilizzo del DW esistente, alto utilizzo prospettico del DW), è quella fase in cui il DW diviene **strategico** per i processi decisionali aziendali. In questa fase si trovano tutte quelle aziende che hanno con successo

intrapreso un progetto di warehousing e che ne stanno sfruttando appieno le potenzialità;

- la quarta fase, chiamata **factory** (alto utilizzo del DW esistente, basso utilizzo prospettico del DW) è la fase in cui si trovano tutte le aziende in cui il DW è maturo, la metodologia di implementazione consolidata e tutte le aree decisionali critiche sono presidiate. In questa fase l'imperativo principale è l'efficienza e il risparmio di costi derivanti dal data warehouse e nel suo utilizzo. Un processo di sclerotizzazione nell'uso del DW può in taluni casi far tornare l'azienda alla prima fase.

1.4.2. Gli ambiti applicativi del DW nelle banche e nelle istituzioni finanziarie

In questo paragrafo individueremo quelle che sono le aree applicative più indicate per il data warehouse nel settore finanziario. Tenteremo anche di ordinare queste aree applicative in funzione della fase della strategia di data warehousing (cfr. paragrafo precedente) in base alla complessità gestita e da gestire.

Controllo di gestione

Può essere questa l'area applicativa di base per un sistema di data warehousing in qualunque organizzazione. In questo caso il data warehouse viene utilizzato sostanzialmente come piattaforma di reporting e analisi di redditività. Come già anticipato nel precedente paragrafo, è inutile e pericoloso ipotizzare di realizzare un data warehouse solo per il controllo di gestione. Tale iniziativa ha senso se, e solo se, questo è il primo passo evolutivo nella strategia di data warehousing dell'azienda. Infatti, costruire un data warehouse per il controllo di gestione consente di analizzare e risolvere rapidamente esigenze estremamente rilevanti ed il cui beneficio è immediatamente chiaro, affrontando peraltro problemi (a livello di struttura, validazione e calcolo dei dati) estremamente noti nella loro struttura.

Risk e Asset Management

Un'altra area applicativa di estremo interesse è identificabile nelle attività di Risk e Asset Management, soprattutto in due attività ben specifiche che sono:

- l'analisi e la simulazione dei portafogli e dei relativi rischi
- il reporting

Tali aree applicative sono di particolare importanza estrategicità ed il data warehouse è lo strumento appropriato per affrontarle, anche per la possibilità di integrare al suo interno dati provenienti anche da fonti esterne all'azienda. E' evidente che in questo caso il data warehouse va dotato di strumentistica di analisi avanzata e basata su algoritmi statistici di analisi e simulazione.

Un'altra sottoarea di grande interesse da questo punto di vista può essere lo sviluppo di sistemi di individuazione delle frodi. Anche in questo caso è necessario il ricorso a strumentazione di tipo statistico per l'implementazione di questo genere di applicazione.

Database di marketing

Non necessariamente il data warehouse è lo strumento appropriato per affrontare e risolvere questo tipo di esigenza. Solo se esiste la necessità di immagazzinare e gestire rilevanti masse di dati probabilmente questa è la scelta giusta. In molti casi il database di marketing è banalmente un'anagrafica clienti arricchita di alcune informazioni "non amministrative" (professione, livello di istruzione, etc.), in casi più avanzati diventa uno strumento fondamentale di supporto al sogno di tutti gli uomini di marketing: il marketing *one-to-one*. In questo caso il database di marketing costituisce una base di informazioni fondamentale per *targettare* correttamente campagne e iniziative promozionali o per attivare servizi avanzati di *customer care.* In questo caso, data la rilevante massa di dati da gestire (p.es.: tutte le operazioni svolte dal cliente), il data warehouse può diventare la piattaforma tecnologica ideale.

Nel settore bancario il marketing one-to-one è ancora allo stadio embrionale, quantomeno dal punto di vista del marketing centrale. Il motivo è semplice: molto spesso il marketing one-to-one viene fatto dalla filiale, l'unica struttura aziendale in grado storicamente di instaurare un rapporto fiduciario con il cliente finale che identifica nello 'sportello' e nel 'suo' impiegato l'azienda.

Supporto Call Center

Anche in questo caso il data warehouse è un'opzione tecnologica, non l'unica praticabile e non necessariamente la più economica. Utilizzare un'architettura di data warehousing a supporto di un'attività di Call Center ha sicuramente senso nel caso in cui le richieste non sono necessariamente di tipo strutturato e quindi risolvibili con il classico "inquiry a terminale". E' evidente però che la tipologia di utente per questo tipo di sistema è più evoluto del normale operatore di *Call Center*.

Un tipico esempio può essere un *Call Center* per una struttura di asset management: la tipica richiesta è la performance composta degli investimenti effettuati dal cliente, un calcolo che richiede accesso a dati provenienti da sistemi informativi potenzialmente diversi e solide competenze di analisi finanziaria.

Knowledge Base

Anche in questo caso valgono le considerazioni fatte per il Database di Marketing: non necessariamente il data warehouse è la tecnologia più idonea per questo tipo di esigenza. Lo è nel momento in cui la conoscenza in oggetto è costituita prevalentemente da informazioni strutturate e preferibilmente numeriche. In tal caso, anche dal punto di vista tecnologico, un database relazionale è sicuramente la soluzione più idonea, efficiente ed economica.

Non è così se invece le informazioni sono di tipo destrutturato (testi, immagini, etc.), nel qual caso probabilmente la soluzione più adatta è una piattaforma di groupware. Si faccia attenzione a non fare confusione con i cosiddetti database multimediali: il fatto che un database relazionale abbia funzionalità multimediali non significa che sia un data warehouse. Infatti, ciò che distingue un data warehouse da ciò che non lo è non è la tecnologia utilizzata, ma l'architettura applicativa ed il disegno della base dati.

Product Engineering

Come già anticipato precedentemente il data warehouse può essere anche una piattaforma decisionale per l'analisi e la concettualizzazione di nuovi prodotti da offrire alla clientela e/o per aggredire nuovi mercati o segmenti di mercato. Tale funzionalità è ovviamente supportata se, e solo se, il data warehouse è dotato non solo di strumenti di analisi dei risultati, ma anche di ambienti di

simulazione che consentono la costruzione ed il testing 'in laboratorio' di nuove soluzioni da proporre ai clienti. In tali ambienti è possibile individuare alcuni importanti aspetti come la marginalità, il punto di pareggio economico, il segmento di clientela interessato, i meccanismi di cannibalizzazione, l'elasticità della domanda e l'impatto sull'equilibrio finanziario aziendale.

Sistema informativo di marketing

E' questo un tipo di applicazione particolarmente interessante. In sostanza si tratta di utilizzare il data warehouse come una sorta di 'backbone' per supportare una serie di applicazioni integrate orientate alle analisi commerciali e di marketing. Gli aspetti fondamentali che caratterizzano questo tipo di architettura sono essenzialmente due:

- la possibilità di integrare basi dati transazionali diverse in un'unica base dati analitica e produrre quindi 'viste' integrate della clientela, del mercato e dei prodotti;
- la possibilità di effettuare analisi con strumenti e logiche diverse su una base dati unica.

L'idea di fondo del sistema informativo di marketing è, infatti, quella che sviluppare un percorso evolutivo che parta dal reporting di base (raccolta, patrimoni, commissioni, etc.) per arrivare ad analisi avanzate (data mining) passando attraverso sistemi di analisi del portafoglio prodotti e clienti e procedure di budgeting e simulazione.

@-business

La diffusione del canale digitale nel settore finanziario pone sicuramente una serie di problemi e di opportunità assolutamente nuove. In primo luogo questo tipo di canale implica una velocità di cambiamento e quindi di reazione nettamente superiore. Il data warehouse può essere lo strumento analitico che consente di cogliere dinamiche all'interno di rilevanti masse di transazioni on-line.

In secondo luogo l'informazione può essere uno strumento si supporto o l'oggetto stesso della transazione ed in questo caso il data warehouse può essere la piattaforma utilizzata per coprire tale ambito applicativo.

Il data warehouse può essere quindi di supporto a sistemi di trading on-line sia dal punto di vista dell'analisi che dal punto di vista dell'architettura dati.

1.5. Considerazioni conclusive

All'interno di questo paragrafo verranno esposte alcune considerazioni conclusive cercando di individuare alcune linee guida per la realizzazione di sistemi di data warehousing. Non esiste in effetti una metodologia unica, consolidata e sempre valida per questo scopo.

Il motivo è semplice: realizzare sistemi diDecision Support non è proprio una scienza esatta, è piuttosto un percorso lungo il quale si effettuano spesso prove e si commettono errori: arrivare al data warehouse 'perfetto' (ammesso che esista) per una specifica azienda è un obiettivo che si ottiene per approssimazioni successive.

Vediamo dunque alcuni 'messaggi' importanti da tenere presente nella realizzazione di un sistema di data warehousing.

1.5.1. Imperativi, dogmi & assiomi

Think big, start small. Non bisogna strafare, ma neanche avere un'ottica di breve. Per essere pratici: non bisogna iniziare un progetto di data warehousing con l'*obiettivo* di realizzare un data warehouse aziendale ma con la *visione* di arrivare a questo traguardo. Cosa significa? Significa che bisogna iniziare da aree analitiche relativamente semplici e conosciute in termini di dati e problematiche e costruire una serie di 'storie di successo' che convincano l'organizzazione che il data warehouse è la strada giusta. Però è sbagliato compiere scelte (soprattutto dal punto di vista tecnologico) che non siano 'scalabili', cioè che richiedano di essere cambiate nel momento in cui la complessità ed il volume dei dati crescono. In sintesi quindi gli obiettivi devono essere a breve, la visione del progetto deve invece guardare necessariamente lontano.

Nel team inserisci solo amanti del data warehouse. Un progetto di data warehouse si può fare solo e unicamente con gente che vive per costruire architetture di warehousing. Il ventaglio delle tematiche da affrontare in tale tipologia di progetto è talmente vasta che per costruire un esperto di data warehousing ci vogliono anni, non mesi. E' questa la ragione per cui chi affronta un progetto di data warehousing deve considerare ciò come l'unico scopo nella propria vita.

Parti dall'esigenza, non dal prodotto. Come già detto, il data warehouse non si 'compra'. Diffidate di chiunque sostenga il contrario, pacchettizzare una piattaforma di decision support non è possibile ed in ogni caso ha delle controindicazioni evidenti. Nello stesso tempo è scorretto selezionare un prodotto e poi costruire il data warehouse con tale prodotto. La *software selection* può derivare unicamente da un'analisi attenta e seria delle esigenze funzionali da soddisfare e dell'infrastruttura tecnica e tecnologica di cui il sistema di data warehousing sarà una componente.

Non banalizzare gli aspetti di business. A chiosa di quanto esposto al punto precedente va rimarcato quest'altro aspetto: la realizzazione di un data warehouse deve assolutamente prendere le mosse da un'analisi degli aspetti di business. Un data warehouse non focalizzato su uno o più specifici aspetti di business sarà un data warehouse senza utenti e quindi destinato a morire di inedia e solitudine.

Non banalizzare gli aspetti tecnologici. Non bisogna però nemmeno eccedere nella direzione esattamente opposta. La realizzazione di un data warehouse pone tali e tanti problemi di *capacity planning* e di definizione architetturale che un'attenzione meno che ossessiva a questo aspetto può decretare il fallimento del sistema. Non sono accettabili tempi di attesa lunghi, né interfacce poco user friendly, né continue *failure* di sistema dovute a mancanza di memoria, CPU, etc. L'analisi dell'infrastruttura tecnologica esistente e il disegno dell'infrastruttura di supporto del data warehouse sono task complessi e fondamentali. Non va dimenticato che di fondo realizzare un data warehouse è un complesso esercizio di system integration.

Educa, cresci e alleva gli utenti. L'interazione con gli utenti è fondamentale sin dalle prime battute del progetto di realizzazione di un data warehouse. Sono gli utenti del sistema che ne devono guidare l'impostazione esplicitando i propri fabbisogni informativi. Sono gli utenti che devono validare l'interfaccia che diventerà la loro 'scrivania' di lavoro. La formazione e la comunicazione sono aspetti fondamentali per questo tipo di progetto.

Attento alla qualità dei dati. Secondo alcune ricerche circa il 50% degli insuccessi nei progetti di data warehousing sono dovuti a problemi di scarsa qualità dei dati alimentanti. E' opportuno, prima di partire con l'implementazione del data warehouse, effettuare una verifica sulla qualità dei dati alimentanti, onde evitare brutte sorprese in seguito. Per la verità molti problemi di data quality vengono scoperti durante il progetto ma un'analisi preliminare ben fatta permette di individuare eventuali aree particolarmente critiche o delle problematiche tipiche dei sistemi informativi alimentanti.

Lascia perdere il report layout. L'approccio tradizionale alla progettazione dei sistemi di supporto decisionale prevede che si parta dai fabbisogni informativi, cioè dalle informazioni di cui gli utenti del sistema vorranno la disponibilità. Per questo motivo moltissima gente, invece di preoccuparsi di analizzare il modello dimensionale di business dell'azienda, si mette a realizzare improbabili e artistici report layout andando in giro a chiedere: "E' questa l'informazione che vuoi vedere ?". Questo tipo di lavoro ha poco senso ed è estremamente pericoloso. Ha poco senso in quanto in molti casi l'utilizzo del data warehouse avviene mediante dei report writer (mediante i quali l'utente stesso può disegnarsi i report) o attraverso degli OLAP browser (in cui il concetto di report non esiste). E' pericoloso perché può capitare che i report finali siano diversi da quelli presentati e inoltre perché da un punto di vista di layout qualsiasi foglio elettronico consente destrutturate perversioni artistiche, inimmaginabili per un *OLAP viewer*, tradizionalmente basato su uno schema dati di tipo rigorosamente multidimensionale.

1.5.2. Gli step di un progetto di data warehousing

Vediamo in breve quali sono i principali step di un progetto di data warehousing.

Analisi delle esigenze funzionali In questa fase devono essere individuati i requisiti funzionali espressi dagli utenti. Tale analisi può essere di fatto ricondotta all'individuazione di specifiche aree applicative e/o informative all'interno dell'architettura.

Disegno logico base dati. In seguito alle analisi effettuate nella fase precedente si procede con il disegno a livello logico della base dati. Tale disegno ha come obiettivo quello di individuare le dimensioni di analisi e le informazioni rilevanti e di ricondurle ad un modello entità-relazioni di riferimento per la prosecuzione del progetto.

Disegno logico flussi alimentanti. Facendo seguito alla fase precedente ed integrandone i risultati con un'attenta analisi dei sistemi alimentanti, devono essere disegnati i flussi di alimentazione del database, sia da un punto di vista di contenuti che modalità organizzative e tempistiche.

Selezione della tecnologia. A questo punto e solo a questo punto è possibile procedere alla selezione della tecnologia di supporto del sistema. Sono chiari infatti sia i volumi da gestire sia le funzionalità richieste dagli utenti per il loro sistema.

Piano di implementazione. In questa fase deve essere disegnato il piano di implementazione del sistema, fase cruciale per dimensionare correttamente le risorse necessarie per l'implementazione. Il piano di implementazione è anche un preziosissimo strumento di coordinamento dei diversi team impegnati poi nel progetto: non è raro infatti che i gruppi di lavoro all'interno del progetto siano almeno tre (interfacce di alimentazione, database e front-end).

Installazione architettura tecnologica. E' la fase in cui l'architettura tecnologica del progetto viene installata. Nel frattempo il gruppo di lavoro può procedere nelle attività di disegno tecnico di dettaglio.

Disegno tecnico di dettaglio base dati e flussi alimentanti. In questa fase devono essere analiticamente progettati i meccanismi di alimentazione della base dati (le query di estrazione dai sistemi

alimentanti, i tracciati record di alimentazione, le regole di trasformazione) e deve essere effettuato il disegno di dettaglio della base dati.

Implementazione base dati. In questa fase si provvede a creare la struttura fisica del database e ad alimentarla con dati di prova per supportare lo sviluppo applicativo.

Implementazione flussi alimentanti. In questa fase vengono sviluppate le interfacce di alimentazione del sistema. Questo stadio è altamente critico in quanto possono venire ad evidenza alcune problematiche di qualità e struttura dei dati alimentanti non individuabili a priori.

Disegno tecnico di dettaglio delle applicazioni. In questa fase le specifiche funzionali della fase analitica vengono tradotte in un disegno di dettaglio che specifica la struttura applicativa in termini di dialoghi, funzionalità e procedure.

Sviluppo applicativo. In questa fase e mediante un approccio di tipo prototipale (cfr. paragrafo seguente) vengono sviluppate le applicazioni facenti parte dell'architettura.

Rilascio e training. A completamento delle attività della precedente fase, vengono rilasciate in produzione le applicazioni e contestualmente viene erogato il training agli utenti.

Come si può comprendere quanto esposto precedentemente è una struttura progettuale di larga massima che contempla le principali attività che devono essere svolte nell'ambito di un progetto di data warehousing. Le attività di dettaglio possono cambiare in maniera sostanziale in funzione dell'ambiente tecnologico, dell'ampiezza del progetto e del tipo di risorse umane disponbili per la sua realizzazione.

1.5.3. L'approccio prototipale

La nostra esperienza dimostra che l'approccio prototipale, ormai peraltro alquanto diffuso nei progetti IT è, anche in questo caso, sicuramente vincente. Il data warehouse è strumento per gli utenti di business che devono essere coinvolti sia nelle fasi progettuali che

nelle fasi implementative del progetto. Ciò per garantire che il prodotto finito sia al 100% 'compliant' con quanto gli utenti di business si aspettano di ottenere.

L'approccio prototipale può essere utilizzato sia per l'intera architettura che per blocchi applicativi facenti parte dell'architettura stessa.

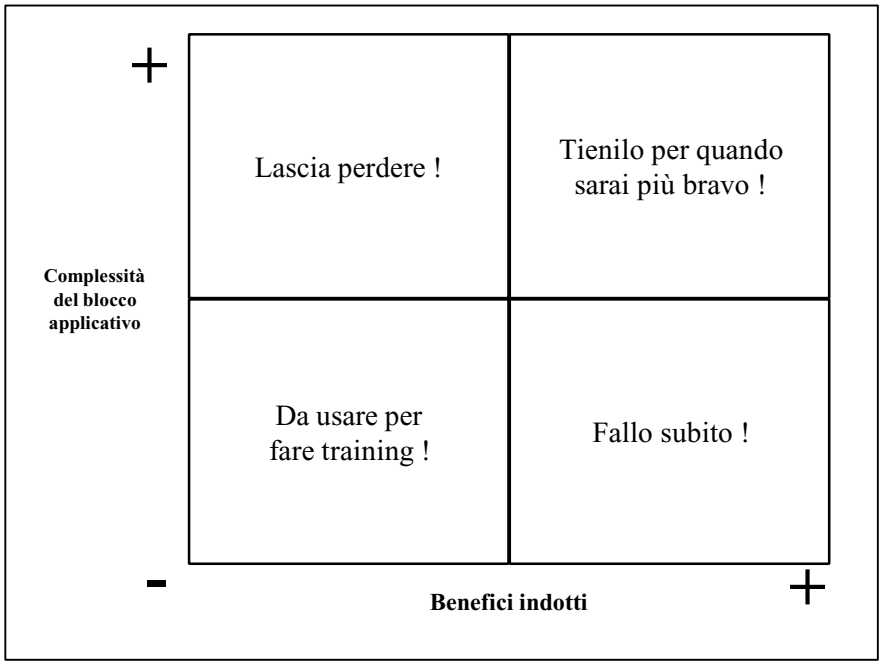


Figura 11 - Approcci alla prototipazione: architettura

Nel primo caso (figura 11) l'approccio migliore consiste nell'identificare un'area analitica a bassa complessità con elevato beneficio per il business o per l'operatività e creare una 'storia di successo' su cui poi proseguire. Eventuali aree a bassa complessità e basso impatto possono essere utilizzate per 'fare training'. Le aree applicative ad alto impatto ed elevata complessità vanno affrontate solo nel momento in cui c'è sufficiente esperienza.

Nel secondo caso (prototipazione applicativa) l'approccio può essere basato su tre iterazioni principali che devono però essere precedute da una fase di disegno tecnico di dettaglio derivante

dall'analisi funzionale effettuata con gli utenti. Le tre iterazioni e le attività collaterali hanno il seguente contenuto.

- **Realizzazione I prototipo (look & feel)**: viene realizzata la prima versione dell'applicazione che presenterà tutte le sue macrofunzionalità. In altre parole saranno realizzati ireport e le relative funzioni di drill down e drill across, le schermate di accesso ai report per l'introduzione dei parametri e le funzioni di calcolo e di elaborazione. Saranno invece tralasciate o solo parzialmente sviluppate le funzioni di utilità, la messaggistica agli utenti, la navigazione tra i report e tra le schermate. Lo scopo principale di questa iterazione è quello di familiarizzare gli utenti con l'interfaccia sia dal punto di vista delle funzionalità sia dal punto di vista 'ergonomico'. La visione di questo primo prototipo verrà effettuata insieme agli sviluppatori.

- **Realizzazione II prototipo (functionality)**: viene realizzata la seconda versione dell'applicazione completa in tutti i suoi aspetti. In altre parole saranno realizzate le modifiche segnalate dagli utenti alla visione del primo prototipo e verranno completate e realizzate tutte le funzionalità ancora mancanti (funzioni di utilità, messaggistica agli utenti, navigazione tra i report e tra le schermate). Contemporaneamente deve essere sviluppata la documentazione utente in maniera tale che gli utenti possano familiarizzare con l'applicazione in modo autonomo ma con il supporto del manuale.

- **Realizzazione applicazione**: viene realizzata la versione definitiva dell'applicazione apportando le modifiche emerse nella revisione precedente. Prima del passaggio in produzione viene effettuato un review generale con gli utenti per l'ultima messa a punto. In seguito a questo step viene aggiornata la documentazione utente in base alle modifiche apportate all'applicazione e viene aggiornato il disegno tecnico dell'applicazione in base alle modifiche apportate nelle precedenti attività di prototipazione.

E' opportuno essere particolarmente attenti, in ogni interazione con gli utenti, a qualsiasi osservazione essi facciano in quanto, molto

spesso, la mancanza di esperienza e di cultura su questo tipo di strumenti possono fare sì che alcune implicazioni sfuggano loro.
Non è raro, infatti, che molte osservazioni e richieste di migliorie arrivino solo dopo diverso tempo dall'entrata in produzione del sistema, dopo che, quindi, gli utenti hanno avuto modo di entrare nelle logiche dello stesso e di familiarizzarvisi.

FONTI BIBLIOGRAFICHE

- V. Poe - *Building a Data Warehouse for Decision Support* - Prentice-Hall - 1996
- R. Kimball - *Data Warehouse Toolkit* - J.W. & Sons - 1996
- AA.VV. - Arthur Andersen Business Consulting - Data Warehousing School Courseware - 1998

Fabio Gasperini, è entrato in Arthur Andersen nel 1986 ed è diventato partner nel 1997. Fa parte della divisione BusinessConsulting di Arthur Andersen MBA nel settore Banche e Finanza. E' responsabile dell'area Integrated Customer Solutions.

Daniele Vanzanelli è entrato in Arthur Andersen nel 1994 ed è diventato dirigente nel 1997. Fa parte della divisione BusinessConsulting di Arthur Andersen MBA. E' responsabile dell'area Data Warehousing & Business Intelligence ed è membro del Data Warehousing Competency Center EMEIA.

WORKSHOP

Data warehousing in ambiente statistico

Aspetti metodologici nella costruzione di un data warehouse statistico

R. Torlone

Aspetti Metodologici nella Costruzione di un Data Warehouse Statistico*

Riccardo Torlone

Dipartimento di Informatica e Automazione
Università di Roma Tre
Via della Vasca Navale, 79 — I-00146 Roma, Italy
E-mail: torlone@dia.uniroma3.it

Sommario

I *data warehouse* sono magazzini storico-temporali di dati statistici di un'organizzazione dedicati ad analisi di supporto a decisioni strategiche da intraprendere. Esistono oggi numerosi strumenti che realizzano molte delle attività connesse alla costruzione e alla gestione di questi magazzini; tuttavia, la tecnologia esistente non è ancora sostenuta da opportune metodologie (se non per attività specifiche) in grado di guidare il progettista nell'intero ciclo di vita di un data warehouse.

In questo lavoro vengono discussi questi aspetti e viene proposta una metodologia di carattere generale, che suggerisce e coordina tutte le attività necessarie alla costruzione e l'uso di un data warehouse, a partire dalla selezione delle sorgenti informative, fino all'organizzazione di collezioni di dati orientate all'analisi multidimensionale. Alcune fasi di questa metodologia corrispondono a tecniche consolidate, per le quali esistono strumenti efficienti dedicati alla loro applicazione; per altre fasi vengono proposti criteri nuovi, basati su modelli di rappresentazione dei dati originali.

1 Introduzione

È ormai ben noto che l'analisi storico-temporale dei dati operazionali offre a tutte le organizzazioni pubbliche e private grandi opportunità per ottenere vantaggi in termini di competitività e di razionalizzazione nell'uso delle risorse. Ad esempio, l'identificazione di andamenti inaspettati nelle vendite può suggerire ad un'azienda commerciale l'opportunità per nuovi affari, mentre l'analisi dei profili d'utenza può essere utile ad una società di servizi per definire tariffe e promozioni. Un *data warehouse* è una collezione integrata e persistente di dati aziendali, orientata al supporto alle decisioni, che viene costruita proprio per favorire queste attività di analisi [4, 6, 7]. Il termine *data warehousing* indica invece il processo di estrazione dei dati dalle varie sorgenti informative aziendali (basi di dati, sistemi legacy, file di vario genere) e di integrazione in un singolo data warehouse.

Un data warehouse è sempre separato fisicamente dalle sorgenti informative, per varie ragioni. Innanzitutto, molto spesso i sistemi operazionali non hanno bisogno di gestire dati storici ed elaborano i dati di dettaglio, mentre i data warehouse devono gestire dati storici, tipicamente in forma aggregata. Inoltre, le varie sorgenti informative possono essere eterogenee, mentre un data warehouse deve offrire una visione integrata dell'intero patrimonio informativo aziendale. Infine, le applicazioni operazionali, generalmente interattive, non devono essere appesantite dalle attività di analisi. Peraltro, i tradizionali sistemi di gestione di basi di dati sono ottimizzati per

*Lavoro finanziato parzialmente dal *CNR* e dal *MURST*.

le attività *transazionali* (On-Line Transaction Processing o OLTP), caratterizzate da molte transazioni concorrenti ognuna delle quali coinvolge spesso pochi record, mentre un data warehouse deve favorire attività *analitiche* (On-Line Analytical Processing o OLAP), caratterizzate da poche interrogazioni, basate su aggregazioni, che coinvolgono molti record [5]. Si tratta di un contesto strettamente connesso a quello delle basi di dati statistiche [10].

In effetti, l'analisi non avviene quasi mai direttamente sul data warehouse, ma piuttosto su speciali collezioni di dati estratte dal warehouse e chiamate *data mart* o basi di dati *multidimensionali*. Il termine multidimensionale deriva dal fatto che l'efficacia dell'analisi è legata anche alla capacità di descrivere e manipolare i dati secondo diverse prospettive chiamate, appunto, "dimensioni." Per esempio, in un'impresa commerciale, le informazioni sulle vendite favoriscono l'analisi orientata al supporto alle decisioni quando sono organizzate secondo dimensioni quali la tipologia di prodotto, il tempo e la località geografica.[1]

Negli ultimi anni si è assistito ad un rapidissimo sviluppo del data warehousing, sia da un punto di vista applicativo che da un punto di vista tecnologico. Esistono oggi molti strumenti in commercio che favoriscono le singole attività e che vengono venduti come moduli di estensione dei sistemi commerciali già esistenti (Oracle, Informix, DB2, SAS, ecc.) oppure come pacchetti applicativi dedicati (ad esempio, Red Brick e Essbase) [9]. Questi strumenti realizzano, dal punto di vista della creazione, attività come il filtering dei dati, il loro caricamento nel data warehouse e l'aggiornamento incrementale. Per quel che riguarda l'uso del warehouse esistono invece i cosiddetti sistemi OLAP, che permettono la specifica di analisi complesse basate tipicamente su aggregazioni lungo le varie dimensioni e la visualizzazione grafica dei risultati dell'analisi. Questi ultimi sistemi si suddividono in sistemi relazionali (ROLAP), che memorizzano i dati in tabelle relazionali, e in sistemi multidimensionali (MOLAP), che rappresentano e memorizzano i dati nella forma di matrici multidimensionali. Recentemente sono emersi anche sistemi "ibridi" che gestiscono entrambi i formati.

Facendo uso di questi, ma anche di strumenti tradizionali, sono state sviluppate negli ultimi tempi moltissime applicazioni reali. Purtroppo però, come spesso avviene, a questo fortissimo sviluppo tecnologico e applicativo non è corrisposto un adeguato sviluppo di metodologie opportune. Oggi la costruzione e la gestione dei data warehouse avviene sulla base di processi empirici, supportati magari da strumenti efficienti e da metodi che favoriscono la soluzione di aspetti specifici, ma mancano del tutto criteri e strategie generali ai quali fare riferimento. Ne consegue che, spesso, la qualità dei prodotti delle analisi è ben al di sotto delle attese ed i costi di creazione e manutenzione dei data warehouse risultano il più delle volte molto elevati rispetto alle previsioni.

In questo lavoro viene proposta una metodologia di carattere generale in grado di guidare il progettista nelle varie attività da compiere lungo *tutto* il ciclo di vita di un data warehouse: dalla selezione delle sorgenti informative, alla costruzione del warehouse, fino alla definizione delle basi di dati multidimensionali. Si seguirà un approccio il piu' possibile generale, sia rispetto alle applicazioni che ai sistemi in gioco, in maniera che la metodologia possa essere utilizzata indipendentemente dal problema allo studio e dagli strumenti a disposizione. Essa è organizzata in un insieme di fasi per alcune delle quali esistono metodologie consolidate (ad esempio, quelle di reverse engineering e di integrazione di schemi) nonché strumenti di supporto; per altre fasi verranno invece proposti modelli e strategie originali.

La metodologia proposta è articolata in quattro "macro-fasi" principali: (1) selezione delle sorgenti informative, (2) integrazione, (3) progettazione del data warehouse e (4) progettazione delle basi di dati multidimensionali. Ognuna di queste fasi è suddivisa in alcuni passi da effettuare

[1] Esiste in effetti un'altra importante modalità di analisi dei dati fattuali che consiste nella ricerca di similarità in transazioni operazionali (il cosiddetto data mining), che comunque non verrà affrontata in questo lavoro.

in cascata, che possono essere a loro volta organizzati in azioni (non necessariamente sequenziali) che si basano su criteri di carattere generale. Verranno usati, per la descrizione dei dati di ingresso e di uscita delle varie fasi, modelli di dati diversi. In particolare, nella prima fase si farà riferimento a modelli logici tradizionali che serviranno a descrivere le sorgenti informative preesistenti. Per la seconda e la terza fase si farà invece riferimento al modello Entità Relazione (o E-R per brevità) secondo la notazione del testo [1], nonché al modello relazionale. Infine, nell'ultima fase si farà riferimento ad un originale modello logico per dati multidimensionali, indipendente dai criteri di rappresentazione usati dai sistemi correnti (tabelle relazionali o matrici dimensionali). Faremo comunque vedere come è possibile implementare schemi del modello logico proposto secondo questi criteri pratici di rappresentazione. In questo modo l'approccio proposto risulta essere generico rispetto agli strumenti, e comunque realizzabile con i prodotti disponibili sul mercato. Il risultato finale offre un inquadramento generale delle attività connesse alla costruzione e all'uso di un data warehouse, che non va ovviamente interpretato come un processo stringente: nei casi pratici, alcune fasi possono essere soppresse perché non necessarie o non realizzabili in pratica, mentre altre possono richiedere attività specifiche ulteriori che non sono riconducibili a criteri di carattere generale. Crediamo comunque che il quadro risultante possa costituire un'utile guida pratica per i progettisti di applicazioni di data warehousing.

Il resto del lavoro è organizzato come segue. Nel Paragrafo 2 verrà presentato l'intera metodologia e verranno brevemente descritti i modelli di dati usati. Nel Paragrafo 3 verranno descritte le prime fasi della metodologia, mentre il Paragrafo 4 è dedicato all'ultima fase relativa alla progettazione della base di dati multidimensionale.

2 La Metodologia e i Modelli Usati

2.1 Presentazione del Quadro Metodologico

La metodologia di riferimento per la costruzione e l'uso di Data Warehouse è illustrato in Figura 1. Essa prevede l'esecuzione di 4 macro-fasi principali, ognuna delle quali è suddivisa in sotto-fasi, con i relativi dati di ingresso e uscita. A loro volta, le varie fasi possono essere suddivise in passi elementari, che suggeriscono azioni da compiere all'interno della specifica attività. Per semplicità, la presentazione delle varie fasi e passi assume che questi siano eseguiti in sequenza; in realtà, come spesso avviene in contesti metodologici complessi, alcune attività possono essere omesse, altre svolte in parallelo, altre infine evidenziare che passi precedenti debbano essere completati o rieseguiti.

La metodologia che si ottiene viene presentata sinteticamente nel seguito ed è preceduta dalla specifica dei dati di ingresso.

Dati di Ingresso. Lo svolgimento della metodologia richiede la specificazione di tre tipologie di informazioni di ingresso: (i) requisiti, (ii) schemi della sorgenti informative aziendali, (iii) schemi di altre sorgenti informativi disponibili. I requisiti sono una descrizione (solitamente in linguaggio naturale semistrutturato) delle esigenze aziendali di analisi. Gli schemi delle sorgenti informative aziendali descrivono formalmente la struttura delle basi di dati operative disponibili, nonché di altre eventuali sorgenti informative (ad esempio, di sistemi legacy), con la relativa documentazione di supporto (tra questa, il glossario aziendale dei termini). Inoltre, spesso accade che l'analisi dei dati aziendali richieda la correlazione di tali dati con altri non di "proprietà" dell'azienda, ma comunque accessibili da essa (ad esempio, statistiche fornite dall'ISTAT, dati sull'andamento della borsa); è allora necessaria una descrizione anche degli "schemi" di tali sorgenti.

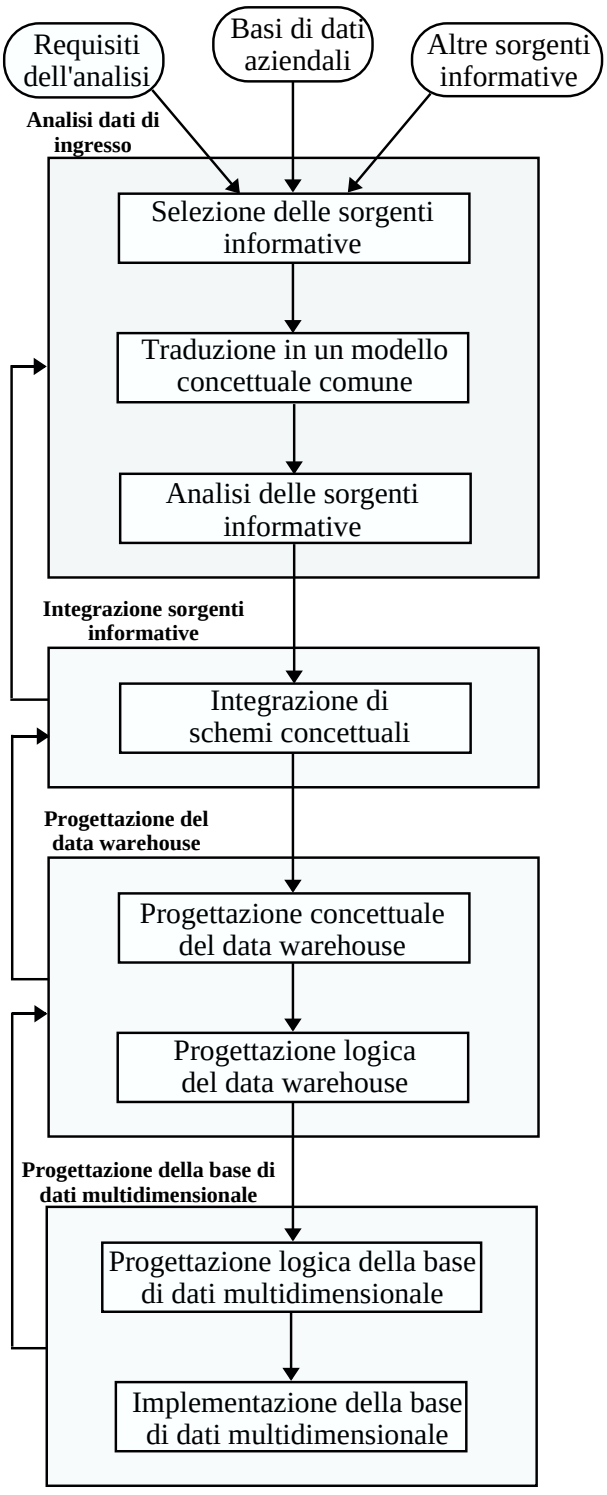


Figura 1: La metodologia di riferimento

Analisi dei Dati di Ingresso. La prima fase consiste in una analisi dei dati a disposizione sulla base dei requisiti dell'applicazione. Viene dapprima effettuata una analisi preliminare del patrimonio informativo, che consiste in una correlazione tra i requisiti e le sorgenti informative disponibili, al fine di selezionare le sorgenti necessarie (o loro porzioni) e individuare eventuali priorità tra schemi. Gli schemi selezionati vengono quindi trasformati in un modello concettuale dei dati di riferimento, per favorire la correlazione tra gli schemi e la loro integrazione. I diversi schemi vengono quindi analizzati al fine di identificare alcuni dei concetti su cui sarà basata l'analisi, nonché criteri per l'integrazione degli schemi.

Integrazione. Le descrizioni concettuali delle sorgenti informative vengono integrate, producendo lo schema concettuale globale del patrimonio informativo aziendale. Questo richiede l'esecuzione dei passi di una metodologia di integrazione, che assumiamo nota. L'integrazione viene guidata dai requisiti, nonché dai criteri di priorità identificati nella fase precedente. Va osservato che lo schema prodotto rappresenta una vista integrata del patrimonio informativo aziendale, ma possiede ancora caratteristiche tipiche delle basi di dati operazionali.

Progettazione del Data Warehouse. Questa fase prevede il progetto e la costruzione del data warehouse. La progettazione concettuale del data warehouse ristruttura lo schema integrato, al fine di introdurre quei concetti necessari per l'analisi che sono ancora assenti nello schema dei dati. La progettazione logica del data warehouse porta alla produzione dello schema logico corrispondente (tipicamente uno schema relazionale), ed elabora la documentazione prodotta al fine di identificare le trasformazioni necessarie alla materializzazione del data warehouse a partire dalle sorgenti informative selezionate.

Progettazione delle Basi di Dati Multidimensionali. Il data warehouse viene costruito per supportare tutte le esigenze aziendali di analisi. In generale, è opportuno estrarre porzioni dei dati del data warehouse, opportunamente aggregate, per realizzare le cosiddette basi di dati multidimensionali (o data mart). Viene dapprima prodotto uno schema multidimensionale per ciascuna esigenza di analisi. In questa fase è opportuno utilizzare un modello logico di dati, in grado di rappresentare esplicitamente quei concetti caratteristici delle basi di dati multidimensionali in maniera indipendente dalle modalità di realizzazione; noi faremo riferimento al modello MultiDimensionale ($\mathcal{MD}$ [3]). Quindi, per ciascuna base di dati multidimensionale viene prodotto il corrispondente schema nel modello dei dati del sistema OLAP a disposizione (relazionale o multidimensionale), nonché definiti i mapping per l'estrazione dei dati dal data warehouse.

2.2 I Modelli di Dati Usati nella Metodologia

Come si evince dalla metodologia proposta, il progettista di un data warehouse ha la necessità di utilizzare una varietà di modelli di dati, in fasi diverse e con finalità diverse.

Innanzitutto, le varie sorgenti informative in ingresso al procedimento di progettazione sono da considerarsi preesistenti, siano esse di proprietà dell'azienda o no. Lo schema di tali sorgenti informative è generalmente disponibile, in un formato che dipende dalla modalità di realizzazione della sorgente stessa. In particolare, se la sorgente è strutturata, allora è descritta da uno schema logico di base di dati (relazionale o non) oppure mediante tracciati record. Altri casi vanno considerati individualmente; ad esempio, è ancora possibile pensare di descrivere dati "semistrutturati" accessibili mediante una interfaccia Web, almeno con riferimento alla loro porzione più regolare. Nei casi più fortunati, è disponibile la documentazione di progetto della sorgente informativa, in cui lo schema dei dati è descritto anche mediante un modello di dati concettuale.

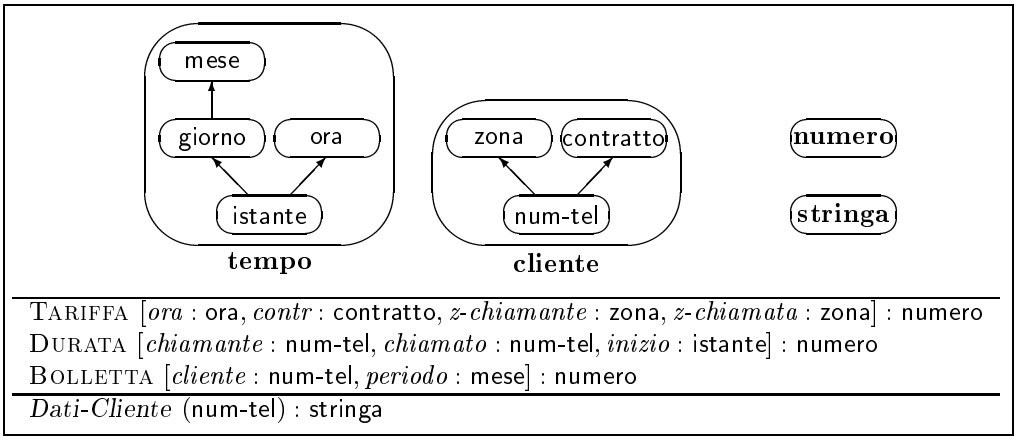


Figura 2: Lo schema $\mathcal{MD}$ `TelCo`

Gran parte della metodologia opera su una descrizione dei dati a livello concettuale. Riteniamo infatti che la manipolazione di più schemi di dati risulti più efficace e più semplice in riferimento ad un modello concettuale, piuttosto che ad un modello logico. In particolare, faremo riferimento al modello Entità Relazione (abbreviato in E-R) [1].

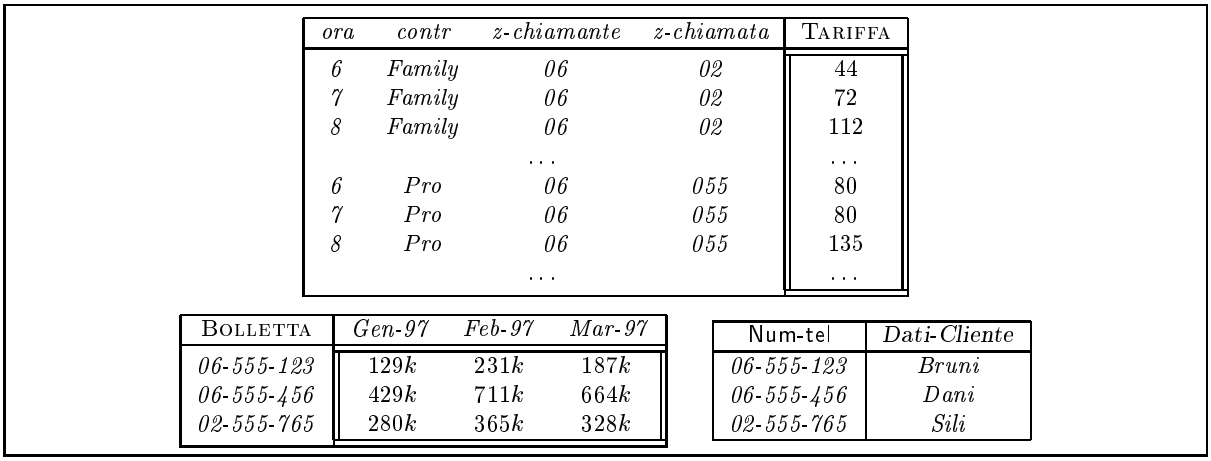
La progettazione del data warehouse viene effettuata prima al livello concettuale (utilizzando ancora il modello E-R) e poi al livello logico. In molti progetti di data warehousing è opportuno realizzare sia un unico data warehouse, che una o più basi di dati multidimensionali. In questi casi, il data warehouse viene spesso opportunamente implementato utilizzando la tecnologia relazionale.

Infine, le basi di dati multidimensionali vengono generalmente implementate utilizzando un server OLAP. Anche in caso di adozione di un server OLAP relazionale, in cui è necessaria una rappresentazione mediante uno schema relazionale, riteniamo opportuno l'utilizzo di un modello di dati per basi di dati multidimensionale che sia indipendente da ogni implementazione. Come già detto, faremo riferimento al modello $\mathcal{MD}$ (descritto nel prossimo paragrafo), che è un modello logico di dati che, nell'ambito dell'analisi multidimensionale, si colloca ad un livello di astrazione maggiore rispetto al modello relazionale. In effetti, nella progettazione di basi di dati multidimensionali, è ragionevole considerare $\mathcal{MD}$ come il modello logico, e il modello relazionale come un modello "fisico," da utilizzare nella fase della progettazione che si occupa dell'allocazione effettiva dei dati e nell'ottimizzazione del loro accesso.

2.3 Il Modello $\mathcal{MD}$

Il modello MultiDimensionale [3] (abbreviato in $\mathcal{MD}$) si basa su due costrutti principali: dimensione e f-tabella. Le *dimensioni* sono categorie sintattiche che consentono di specificare "prospettive" secondo le quali vogliamo analizzare i dati. Ogni dimensione è organizzata in una gerarchia di *livelli*, che corrispondono essenzialmente a domini a livelli di granularità differenti. È possibile associare *descrizioni* ai livelli. In una dimensione, valori di livelli differenti sono correlati da una famiglia di *funzioni di roll-up*. Le *f-tabelle* sono funzioni che associano *misure* a *coordinate simboliche* (definite rispetto a una particolare combinazione di livelli): esse sono usate per rappresentare i dati fattuali.[2]

[2]La "f" nel termine f-tabella ha un doppio significato: sta per "funzione," perché ogni f-tabella è in effetti una funzione da coordinate a misure; e per "fatto," perché rappresenta ciò che nei sistemi è comunemente memorizzato nelle "tabelle dei fatti" (fact-tables).



ora	*contr*	*z-chiamante*	*z-chiamata*	TARIFFA
6	*Family*	*06*	*02*	44
7	*Family*	*06*	*02*	72
8	*Family*	*06*	*02*	112
		...		...
6	*Pro*	*06*	*055*	80
7	*Pro*	*06*	*055*	80
8	*Pro*	*06*	*055*	135
		...		...

BOLLETTA	*Gen-97*	*Feb-97*	*Mar-97*
06-555-123	129*k*	231*k*	187*k*
06-555-456	429*k*	711*k*	664*k*
02-555-765	280*k*	365*k*	328*k*

Num-tel	*Dati-Cliente*
06-555-123	*Bruni*
06-555-456	*Dani*
02-555-765	*Sili*

Figura 3: Una istanza dello schema `TelCo`

Consideriamo ad esempio una compagnia telefonica interessata all'analisi dei suoi dati operazionali. I dati relativi alle chiamate possono essere organizzate lungo le dimensioni **tempo** e **cliente**. Le corrispondenti gerarchie sono riportate in Figura 2. Le altre dimensioni sono usate per rappresentare valori numerici e stringhe (dimensioni atomiche **numero** e **stringa**). Il livello num-tel (numero telefonico) si aggrega su zona (l'area geografica dell'utenza telefonica, identificato da un codice) e contratto (caratterizzato da tariffe in ore differenti). Il dominio associato al livello istante contiene "timestamps" del tipo *10:45:21::5-Gen-97*. Questo valore si aggrega in *10* al livello di ora (inteso nel senso di "ora del giorno") e in *5 Gen 97* a livello giorno. Diverse f-tabelle possono essere definite in questo contesto (come indicato nella stessa figura). TARIFFA rappresenta il costo per un minuto di conversazione tra un cliente nella zona *z-chiamante* (con un contratto di tipo *contr*) e un cliente nella zona *z-chiamata*, con inizio in una certa *ora*. La seconda f-tabella associa la DURATA in secondi ad ogni chiamata (fatta da un *chiamante* a un *chiamato* ad una certa ora). La f-tabella BOLLETTA è derivata e aggrega i costi per numero di telefono e mese. Infine, *Dati-Cliente* è una descrizione di livello che associa ad una utenza telefonica i dati anagrafici del cliente.

Una possibile istanza per lo schema `TelCo` viene riportato in Figura 3. Una coordinata simbolica sulla f-tabella TARIFFA è

$$[ora : 7, contr : Family, z\text{-}chiamante : 06, z\text{-}chiamata : 02].$$

L'istanza associa la misura 72 a questa entry. La descrizione *Dati-Cliente* associa la stringa *Bruni* al valore *06-555-123* del livello num-tel. Si osservi che sono state usate in Figura 3 due rappresentazioni grafiche diverse per una f-tabella: una tabella tradizionale per TARIFFA e una matrice per BOLLETTA. Questo fatto suggerisce che è in effetti possibile implementare una f-tabella in diverse maniere.

3 Fasi per la Progettazione del Data Warehouse

La progettazione di un data warehouse richiede l'esecuzione delle prime tre macro-fasi tra le quattro previste dalla metodologia. Tra queste, le prime due sono prevalentemente composte da attività descritte da metodologie note in letteratura. In effetti, la progettazione di un data warehouse ha molte caratteristiche in comune con la progettazione di una base di dati integrata

o federata, tranne che per le finalità, decisamente rivolte alla gestione di dati per l'analisi anziché ad attività transazionali. Nei Paragrafi 3.1 e 3.2 vengono illustrate brevemente le attività preliminari che portano essenzialmente alla definizione di uno schema concettuale integrato del patrimonio informativo aziendale. Il Paragrafo 3.3 è dedicato alla descrizione delle attività da svolgere per progettare il data warehouse (orientato all'analisi) a partire dallo schema concettuale integrato (che ha prevalentemente caratteristiche "transazionali").

3.1 Analisi dei Dati di Ingresso

La prima fase consiste in una attenta analisi dei dati a disposizione sulla base dei requisiti dell'applicazione e può essere suddivisa nelle seguenti sotto-fasi.

Selezione delle Sorgenti Informative. L'analisi preliminare del patrimonio informativo disponibile consiste in una sua correlazione con i requisiti. Alcune sorgenti informative (o loro porzioni) risulteranno irrilevanti ai fini dell'analisi, e verranno per questo motivo trascurate. Inoltre è possibile assegnare delle preferenze tra gli schemi; ad esempio, in riferimento alla rappresentazione di un certo concetto, sarà preferibile lo schema della sorgente informativa in cui tale concetto è gestito in modo più accurato e aggiornato.

Traduzione in un Modello Concettuale di Riferimento. Questa fase prevede la traduzione degli schemi delle sorgenti informative selezionate in un modello dei dati comune, di riferimento. Come già accennato, utilizzeremo come modello di dati comune un modello concettuale, ed in particolare il modello E-R. Riteniamo infatti che la correlazione di più schemi di dati sia più semplice ed efficace in riferimento a un modello concettuale piuttosto che ad un modello logico. Il prodotto è costituito da uno schema E-R per ciascuna sorgente informativa selezionata, corredato dalla opportuna documentazione di supporto (ad esempio, metadati da utilizzare per l'eventuale conversione di dati fattuali).

Analisi delle Sorgenti Informative. Questa fase, preliminare all'integrazione di schemi, ha lo scopo di individuare, nei vari schemi, concetti irrilevanti o significativi per l'analisi. Sono da considerarsi irrilevanti quei concetti presenti nello schema per motivi strettamente operativi; in questo caso non ci aspettiamo che l'analisi si concentri su tali dati. I corrispondenti concetti possono essere rimossi dai rispettivi schemi. D'altra parte, è importante identificare i concetti candidati alla rappresentazione di "fatti," "misure" e "dimensioni" nelle specifiche attività di analisi orientata al supporto alle decisioni. In questo contesto, chiamiamo *fatti* i concetti dello schema E-R (entità, relazioni o attributi) sui quali il processo di analisi è centrato. Una *misura* è invece una proprietà atomica di un fatto che intendiamo analizzare (tipicamente un attributo numerico di un fatto o un conteggio delle sue istanze). Infine, una *dimensione* è un sottoinsieme dello schema E-R dato che descrive una prospettiva lungo la quale l'attività di analisi può essere effettuata. Inoltre, per concetti di questi tipi che risultano essere rappresentati in più schemi, è necessario identificare eventuali sovrapposizioni delle istanze corrispondenti e, in questo caso, delle priorità.

3.2 Integrazione

Le descrizioni concettuali delle sorgenti informative vengono integrate, producendo lo schema concettuale globale del patrimonio informativo aziendale. L'integrazione di schemi concettuali è stata largamente studiata e può essere effettuata secondo i criteri descritti, per esempio, in [2]. Citiamo il fatto che questa attività può essere utilmente guidata dai requisiti, nonchè dai criteri di priorità identificati nella fase precedente. (Ad esempio, la risoluzione di eventuali

conflitti strutturali può essere guidata dal criterio che privilegia la rappresentazione di "fatti" come entità.) Viene inoltre prodotta la necessaria documentazione di supporto orientata alla integrazione dei dati.

3.3 Progettazione del Data Warehouse

Questa fase prevede il progetto e la costruzione del data warehouse ed è vantaggioso suddividerla in due passi distinti.

Progettazione Concettuale del Data Warehouse. Questa fase porta alla produzione dello schema concettuale del data warehouse a partire dallo schema integrato delle sorgenti informative a disposizione. Infatti, lo schema integrato prodotto dal passo precedente contiene i concetti necessari per le esigenze di analisi, ma con caratteristiche che sono ancora quelle di dati rappresentati per esigenze operative. È quindi necessario introdurre nello schema degli ulteriori concetti necessari per l'analisi, che sono spesso complementari a concetti già presenti nello schema. In questa fase devono essere aggiunti, ad esempio, dati dimensionali ancora assenti (in particolare, concetti per la rappresentazione di dati storici) e dati aggregati. È possibile inoltre rimuovere concetti irrilevanti per l'analisi che non sono stati precedentemente identificati. Il prodotto di questa fase è lo schema concettuale del data warehouse, con la documentazione di supporto che descrive anche le procedure di estrazione e integrazione delle istanze dalle diverse sorgenti informative.

Progettazione Logica del Data Warehouse. A partire dallo schema concettuale del data warehouse, ne viene progettato il corrispondente schema logico (tipicamente, nel modello relazionale). Questa attività può essere svolta utilizzando metodologie note (vedi ancora [1]). I criteri di progettazione da utilizzare sono talvolta diversi da quelli tradizionali: infatti il data warehouse è spesso caratterizzato da un opportuno livello di denormalizzazione dei dati, con presenza di ridondanza, soprattutto per rappresentare dati derivati.

Deve essere inoltre prodotta la documentazione di supporto che descrive la sequenza di trasformazioni necessarie all'estrazione e all'integrazione dei dati dalle sorgenti informative al data warehouse. In particolare, questa deve descrivere sia i mapping da utilizzare per popolare inizialmente il data warehouse, che quelli necessari a propagare gli aggiornamenti.

4 Progettazione delle Basi di Dati Multidimensionali

Come osservato nell'Introduzione, l'analisi dei dati non avviene direttamente sul data warehouse, ma piuttosto nei data mart, ovvero in basi di dati multidimensionali derivate dal data warehouse. In effetti, i data mart sono spesso descritti in letteratura come dei data warehouse "dipartimentali," in quanto sono orientati al supporto di singole esigenze di analisi. Tuttavia, i dati del data mart sono semplicemente estratti dal data warehouse già costruito, anziché estratti e integrati dalle singole sorgenti informative operazionali.

Questo paragrafo descrive la fase di progettazione di un data mart, a partire dallo schema concettuale del data warehouse ed in riferimento ad una singola esigenza di analisi. In particolare, la nostra metodologia sarà illustrata in riferimento allo schema concettuale del data warehouse `Rivendita`, illustrata in Figura 4, che possiamo supporre costruito utilizzando le fasi metodologiche descritte nel Paragrafo 3.

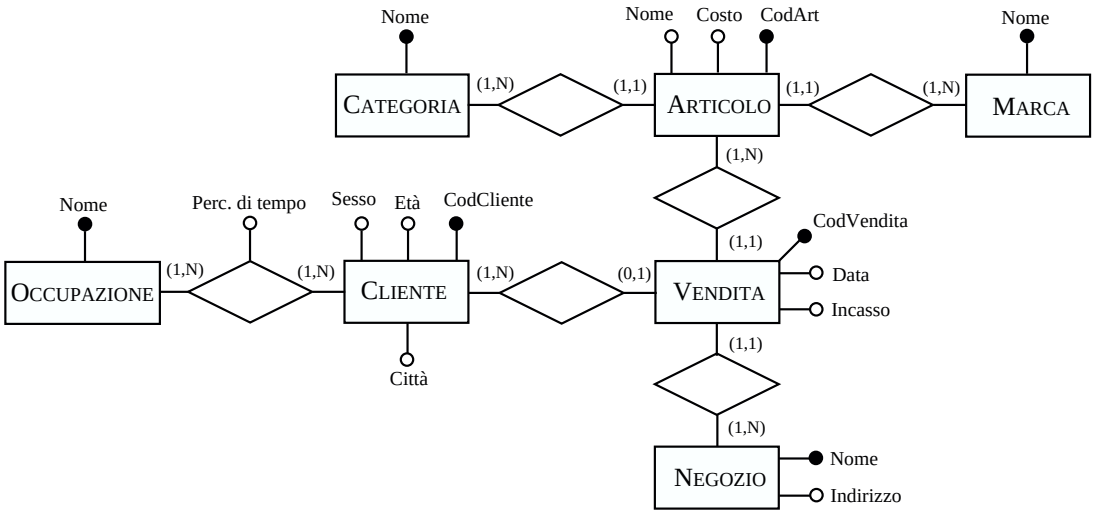


Figura 4: Lo schema E-R che descrive il data warehouse `Rivendita`

4.1 Progettazione Logica di una Base di Dati Multidimensionale

La fase di progettazione logica di una base di dati multidimensionale viene suddivisa in quattro passi: (i) identificazione di fatti e dimensioni; (ii) ristrutturazione dello schema E-R; (iii) derivazione di un "grafo dimensionale"; (iv) traduzione nel modello $\mathcal{MD}$. In effetti, i primi due passi non sono strettamente sequenziali, ma procedono, in molti casi, in parallelo: durante la ristrutturazione dello schema E-R, i fatti e le dimensioni selezionati possono essere raffinati e/o modificati. Successivamente, il processo procede sequenzialmente, poiché ogni passo richiede il completamento del passo precedente.

Senza entrare nei dettagli di questa fase (il lettore interessato può consultare il lavoro [3]), lo schema E-R della nostra applicazione che si ottiene dopo la fase di ristrutturazione è riportato in Figura 5. Si osservi che lo schema è stato annotato con fatti e dimensioni. Si osservi inoltre che le dimensioni non contengono gli attributi descrittivi (ad esempio, l'attributo *Indirizzo* dell'entità NEGOZIO).

4.1.1 Derivazione di un Grafo Dimensionale.

Partendo dallo schema E-R ristrutturato, possiamo ora derivare un grafo speciale che chiameremo *dimensionale*. Un grafo dimensionale rappresenta, in maniera succinta e facilmente comprensibile, fatti e dimensioni dello schema E-R di partenza. La Figura 6 riporta il grafo dimensionale che si ottiene dallo schema E-R di Figura 5. È facile comprendere che questo grafo può essere ottenuto automaticamente e che possiede lo stesso contenuto informativo dello schema E-R originale. Si osservi inoltre che le dimensioni diventano sottografi del grafo dimensionale. In un grafo dimensionale è possibile distinguere quattro tipi di nodi: *nodi fatto*, denotati da margini in grassetto (originano da entità fatto); *nodi livello*: sono quelli contenuti in una dimensione; *nodi descrittivi*: sono nodi non contenuti in dimensioni e hanno un arco entrante che esce da un nodo livello (originano da attributi descrittivi); *nodi misura*: sono nodi non contenuti in dimensioni e hanno un arco entrante che esce da un nodo fatto (essi originano da misure).

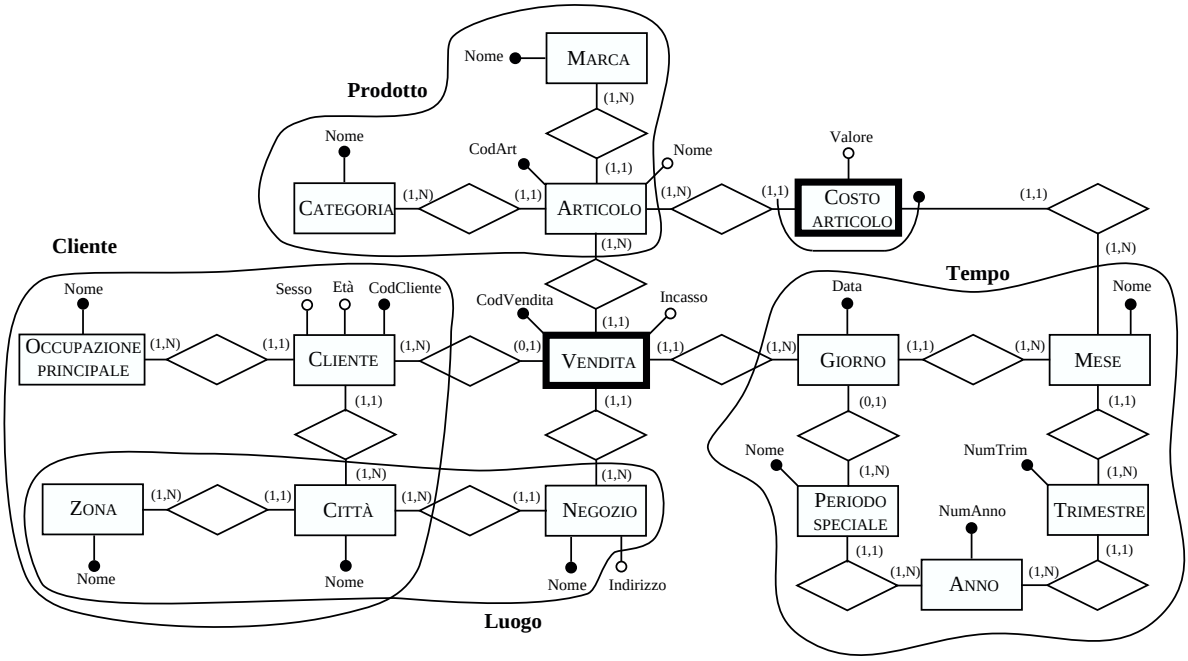


Figura 5: Lo schema E-R di Figura 4 dopo la fase di ristrutturazione

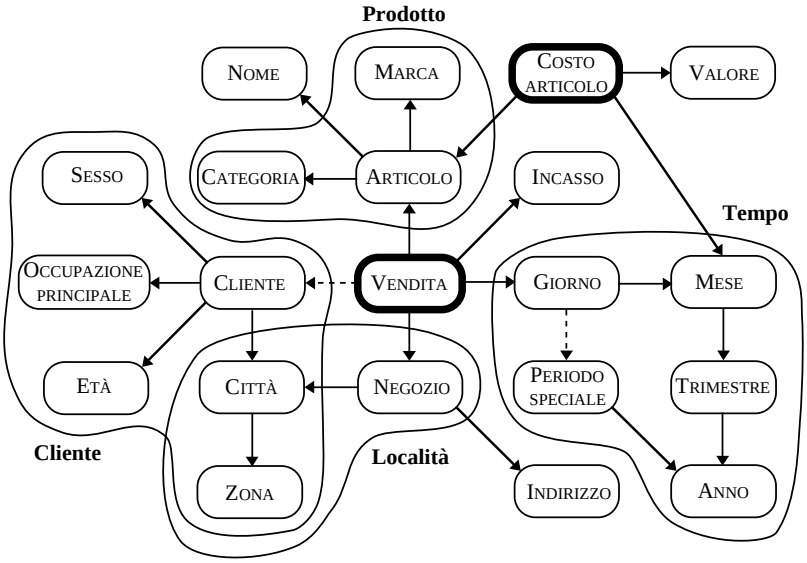


Figura 6: Il grafo dimensionale ottenuto dallo schema in Figura 5

4.1.2 Traduzione nel Modello $\mathcal{MD}$.

Da un grafo dimensionale è possibile ottenere direttamente le $\mathcal{MD}$-dimensioni: ce ne sarà una per ogni dimensione del grafo e, per ognuna di esse, avremo un $\mathcal{MD}$-livello per ogni nodo e una funzione di roll-up per ogni arco del corrispondente sottografo. Tali sottografi denotano anche l'ordinamento parziale sugli $\mathcal{MD}$-livelli.

Le f-tabelle possono essere definite come segue. Per ogni nodo fatto del grafo dimensionale, selezioniamo una combinazione di livelli appartenenti alle dimensioni "associate" al fatto, quelle cioè per le quali esiste un arco uscente dal nodo fatto che incide su di esse (in particolare sul loro livello base). Per ogni dimensione si possono selezionare più livelli per ciascuna dimensione e non è necessario selezionare un livello per ogni dimensione associata al fatto. Dobbiamo quindi definire un mapping θ, basato eventualmente su aggregazioni, che descriva il risultato della f-tabella. Questo mapping può essere (i) il conteggio di una collezione di fatti, oppure (ii) una espressione su una misura. Le istanze delle f-tabelle sono poi definite di conseguenza [3].

Nell'applicazione `Rivendita`, avevamo identificato tre misure: (1) il numero di articoli venduti, (2) gli incassi e (3) il costo degli articoli. Le prime due misure vanno descritte giornalmente per ogni articolo e negozio, la terza va descritta su base mensile. Queste misure possono essere descritte dalle seguenti f-tabelle.

1. VENDITA[*periodo* : giorno, *prodotto* : articolo, *luogo* : negozio] : numerico, definita sul fatto VENDITA dal mapping **count**(VENDITA);

2. INCASSOTOT[*periodo* : giorno, *prodotto* : articolo, *luogo* : negozio] : numerico, definita sul fatto VENDITA dal mapping **sum**(INCASSO(VENDITA));

3. COSTOARTICOLO[*periodo* : mese, *prodotto* : articolo] : numerico, definita sul fatto COSTO ARTICOLO dal mapping VALORE(COSTO ARTICOLO).

Potremmo essere interessati anche a dati parzialmente aggregati. Ad esempio, l'analisi degli acquisti dei clienti per età, categoria di prodotto e anno può essere effettuata con la seguente f-tabella:

$$\text{VENDITEPERETA}[età : \text{età}, prodotti : \text{categoria}, periodo : \text{anno}] : \text{numerico},$$

definita sul fatto VENDITA dal mapping **sum**(INCASSO(VENDITA)).

4.2 Implementazione di una Base di Dati Multidimensionale

Una base di dati multidimensionale può essere facilmente implementata mediante un sistema OLAP. Per problemi di spazio verrà brevemente illustrata solo l'implementazione mediante tabelle relazionali. Una descrizione più dettagliata dell'implementazione su sistemi OLAP relazionali e multidimensionali è presentata in [3].

La rappresentazione naturale di una base di dati multidimensionale nel modello relazionale consiste in uno "schema a stella" (*star scheme* [6, 7]) contenente "tabelle fatto" e "tabelle dimensione". Le prime sono normalizzate mentre le seconde possono essere denormalizzate. Lo schema stella rappresentante una base di dati $\mathcal{MD}$ può essere costruito come segue. Abbiamo: (1) uno schema di relazione R_d per ogni dimensione non-atomica d, e (2) uno schema di relazione R_f per ogni f-tabella f. Le dimensione atomiche non vanno rappresentate esplicitamente dato che generalmente corrispondono a domini di base. Ad esempio, la rappresentazione a schema stella della base di dati $\mathcal{MD}$ `Rivendita` è mostrata in Figura 7 (l'istanza è incompleta).

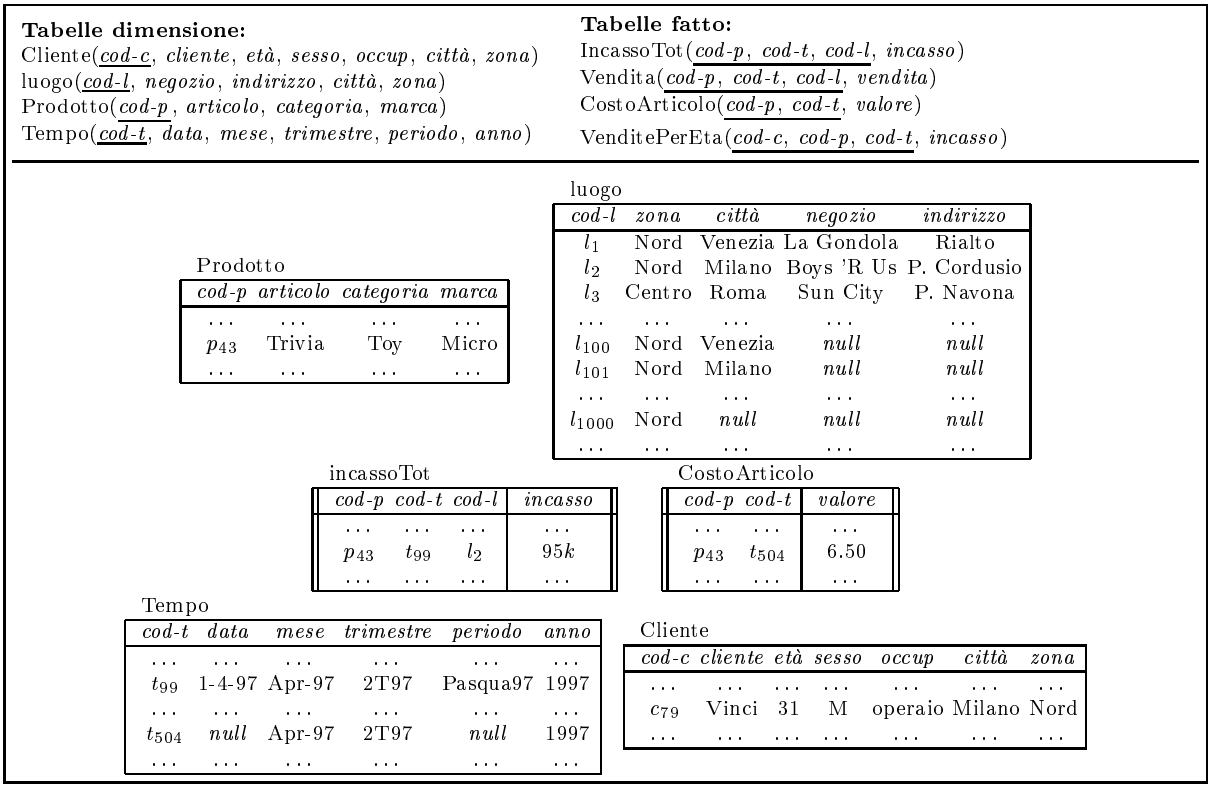


Figura 7: Lo schema a stella della base di dati multidimensionale `Rivendita`

Bibliografia

[1] C. Batini, S. Ceri, and S. Navathe. *Conceptual Database Design*. Benjamin/Cummings, 1992.

[2] C. Batini, M. Lenzerini, and S.B. Navathe. A comparative analysis of methodologies for database scheme integration. *ACM Computing Surveys*, 18(4):323–364, 1986.

[3] L. Cabibbo and R. Torlone. A logical approach to multidimensional databases. In *Int. Conf. on Extending Data Base Technology (EDBT'98), Springer-Verlag*, 1998.

[4] S. Chaudhuri and U. Dayal. An overview of Data Warehousing and OLAP technology. *ACM SIGMOD Record*, 26(1):65–74, March 1997.

[5] E.F. Codd, S.B. Codd, and C.T. Salley. Providing OLAP (On Line Analytical Processing) to user-analysts: An IT mandate. Arbor Software White Paper, *http://www.arborsoft.com*.

[6] W.H. Inmon. *Building the Data Warehouse*. John Wiley & Sons, second edition, 1996.

[7] R. Kimball. *The Data Warehouse toolkit*. John Wiley & Sons, 1996.

[8] A. O. Mendelzon. Data warehousing and OLAP: a research-oriented bibliography. *http://www.cs.toronto.edu/~mendel/dwbib.html*.

[9] N. Pendse and R. Creeth. The OLAP Report. *http://www.olapreport.com*.

[10] A. Shoshani. OLAP and statistical databases: Similarities and differences. In *Sixteenth ACM SIGACT SIGMOD SIGART Symp. on Principles of Database Systems*, pages 185–196, 1997.

WORKSHOP

Data warehousing in ambiente statistico

SAS

W. Lanzani

Il Data Warehouse in ambiente statistico

SAS

La realizzazione di una banca dati sia pubblica che privata, sia di dimensioni limitate che estesa e complessa, è sempre un lavoro impegnativo che, se viene supportato da strumenti adeguati, può trasformarsi in un vero e proprio inferno. E' altresì vero che, in un mondo complesso come quelle in cui stiamo vivendo, l'informazione, di qualità e tempestiva, sta diventando sempre più un elemento chiave del successo, capace di migliorare le prestazioni e diminuire i costi.

Va anche ricordato come gli enti pubblici per poter affrontare la rivoluzione organizzative e gestionale derivante dalla riforma di cui sono protagonisti, utilizzano sempre più metodologie tipiche del settore privato, come il controllo della qualità e la regolamentazione dei contratti. Anche le numerose privatizzazioni fatte in questi anni hanno portato gli enti pubblici interessati a fronteggiare per la prima volta un ambiente di mercato competitivo.

Dal mix di questi elementi, necessità di informazioni e adozione di mentalità proveniente dal settore privato, risulta immediatamente chiaro come sia di basilare importanza poter contare, sempre, su un sistema di supporto decisionale affidabile.

SAS è sicuramente una delle società più attive ed impegnate nello sviluppo di soluzioni software espressamente pensate per la Pubblica Amministrazione.

La strategia SAS è incentrata sull'Information Delivery, cioè sulla capacità di fornire alla persona giusta, nel momento giusto, qualunque informazione di cui abbia bisogno. Questa è la sfida attuale dei sistemi informativi, sfida prima raccolta dal settore privato ora indirizzata anche al settore pubblico. Per poter affrontare al meglio questa sfida bisogna strutturarsi con sistemi di Data Warehouse, necessario per organizzare i dati permettendo analisi rapide senza ulteriori preoccupazioni per gli utenti finali.

Il Data Warehouse è in grado di raccogliere enormi quantità di dati proveniente da diverse fonti, ripulirli, catalogarli, e riconsegnarli alla persona giusta al momento giusto. Il tutto nella forma aggregata secondo le specifiche esigenze in termini statistici o analitici e consentendo di trarre conoscenza da queste informazioni mettendole in relazioni appropriata, operando comparazioni storiche, geografiche e di qualunque altra natura. Fino all previsione di modelli comportamentali di fenomeni complessi e allo studio degli stessi in relazione a processi e obiettivi prestabiliti, per arrivare a vere e proprie strutture informative basate su logiche di 'artificial intelligence'.

SAS è in grado di offrire potenti strumenti che consentono la gestione e l'analisi dei dati, esplorabili sotto un numero praticamente illimitato di prospettive; questo grazie alle tecniche di Business Intelligence quali OLAP, analisi multidimensionale, data mining, EIS, sfruttando anche le tecnologie offerte da Internet, ma tutte queste tecniche si basano sulle solide fondamenta di un buon Data Warehouse.

Certamente l'analisi statistica è una fra le branche dello studio dei dati che appare come un settore quanto mai dedicato all'attenzione di pochi addetti del settore, ma in alcuni progetti si nota sempre di più la tendenza a presentare l'output dell'analisi in maniera sintetica e comprensibile da parte di chi addetto ai lavori non è, fornendo strumenti di front end user friendly che permettono di interagire con il sistema di analisi eventualmente para metrizzandolo e ricevendone l'output nella forma più appropriata.

Alcuni casi di data warehouse statistico

Regione Piemonte

Un esempio di Data Warehouse statistico è sicuramente quello della Regione Piemonte che in collaborazione con il CSI-Piemonte e SAS ha realizzato il progetti BDDE (Banca Dati Demografica Evolutiva).

L'obiettivo di questo progetto era quello di realizzare un archivio, a livello regionale, in grado di fornire un'informazione demografica completa, aggiornata e dettagliata, generata attraverso il coordinamento e l'integrazione di svariati flussi di dati, spesso non coincidenti e provenienti da fonte diverse. La realizzazione di un simile progetto,

elemento fondamentale nel supporto alle decisioni di programmazione, è stata resa particolarmente ardua dalle caratteristiche di elevata frammentazione amministrativa che caratterizzano la regione piemontese; la collaborazione traCSI-Piemente, uffici regionali ed Università ha portato alla progettazione di un Data Warehouse in ambiente SAS. Tra gli obiettivi raggiunti nell'attuale implementazione della BDDE emergono sicuramente: l'articolazione dei dati demografici in vari livelli di dettaglio (geografici, per fascia d'età e per sesso) la possibilità di navigare fra i vari flussi ottenendone dati, indicatori e rappresentazioni grafiche; la certificazione dei livelli di qualità dei dati raggiunti tramite continue verifiche e controlli; l'aggiornamento continuo dei dati stimati all'arrivo dei dati definitivi, la facile integrazione nella struttura originaria di moduli aggiuntivi che ne espandono la capacità; la costruzione di indici sintetici di condizione e sviluppo demografico ed infine l'accessibilità alla BDDE da parte di svariate tipologie di utenza, adattando quindi i dati forniti alle esigenze degli utilizzatori. I dati inseriti nel database costituiscono un archivio storico della popolazione piemontese i cui dati iniziano dal 1991 e sono alimentati dalle statistiche ufficiali ISTAT. Ogni volta che vengono aggiornati i dati, il sistema crea delle stime sull'andamento della popolazione basate sui dati precedenti, l'utilizzo delle stime da parte degli amministratori consente di avere a disposizione dati statistici di ottima qualità e molto vicini a quelli definitivi. Le stime sono generate da un'equazione di contabilità demografica che è stata migliorata nel corso degli anni; il sistema attuale è in grado di fornire una previsione la cui attendibilità è relazionata alla data dell'ultimo dato certo. I risultati ottenuti dal progetto hanno portato ad una proficua collaborazione con gli uffici ISTAT di Torino e Roma, culminata con la firma di una convenzione tra ISTAT e Regione Piemonte che delega agli uffici regionali le rilevazioni POSAS:
Inoltre, la Regione Piemonte utilizza la BDDE anche in ambito sanitario; i principali processi sono stati realizzati con le tecnologie SAS, con l'obiettivo di integrare i dati provenienti da ISTAT e Data Warehouse regionale sulla mortalità per causa all'interno del sistema.
Anche in questo caso si è cercato di realizzare uno strumento che consenta un accesso alle informazioni più facile, in modo rapido e comprensibile. Per raggiungere questi obiettivi tutte le informazioni contenute nella banca dati sono visualizzabili tramite chiari grafici a barre, è inoltre possibile effettuare il confronto fra i dati utilizzando un ampio numero di disaggregazioni. I dati, oltre che per età e sesso, sono disaggregabili per provincia e ASL

ISTAT (relazione presentata in occasione del Convegno SUGItalia '2000)

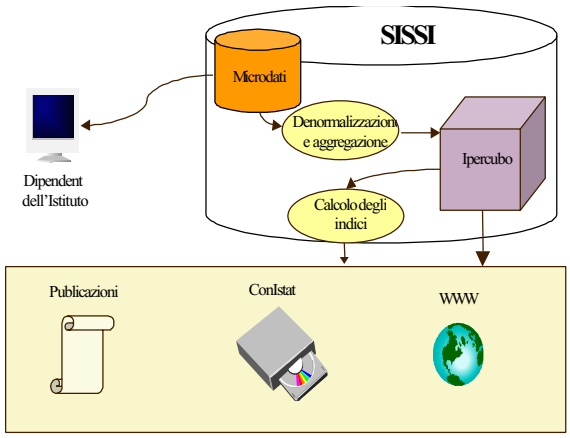
Abstract
Il ***Data Warehouse*** è il naturale complemento del tradizionale Data Base. Passare dal dato all'informazione per arrivare alla comprensione dei fenomeni gestionali è vitale per il*management* di qualsiasi organizzazione.
L'accrescersi di importanza del Web comporta tuttavia la revisione di alcuni dei criteri sui quali si fondavano i Data Warehouse del passato, si parla infatti di Data WebHouse per sottolineare la possibilità di avere informazioni di un warehouse mediante la rete Internet/Intranet.
Il WebHouse SISSI (Sistema Informativo Statistiche Strutturali d'Impresa) nasce dall'esigenza di integrare le informazioni sull'impresa derivanti dalle diverse indagini statistiche, e di distribuire le informazioni raccolte via web.
SISSI è un progetto che, iniziato come semplice data&informationbase, è sfociato in un sistema più complesso contenente da un lato le enormi quantità di dati, dall'altro le applicazioni per la loro diffusione su Internet.
Con l'utilizzo della tecnologia SAS si è realizzata un'unica fonte di dati certificati, fruibili via web mediante semplici browser, senza alcun tipo di installazione ulteriore sui PC degli utenti.
I moduli SAS utilizzati per la realizzazione dell'intero processoend-to-end sono:

- SAS/Warehouse Administrator per l'analisi, l'integrazione e l'aggregazione dei dati confluenti nel warehouse e la creazione di un data Mart relativo a una particolare esigenza utente;
- Sas/IntrNet & AppDevStudio per la diffusione di report, tavole statistiche, query dinamiche su Web.

L'applicazione di WebHouse, per il momento disponibile solo per gli utenti interni, offre la libertà e la possibilità di analizzare i dati in modi diversi, senza essere vincolati ad una serie di prospetti predefiniti.

Introduzione
SISSI è il Sistema Informativo Statistico Sulle Imprese, sviluppato all'interno della Direzione Centrale delle statistiche su Istituzioni e Imprese

(DCII) dell'Istat. Esso rappresenta una struttura multidimensionale e multifunzionale che tende a coprire tutte le informazioni prodotte dalla DCII con riferimento alle imprese.
SISSI fa parte di un sistema più ampio, il Sistema Informativo delle Statistiche sulle Imprese E le Istituzioni (SISSIEI), che comprende anche le informazioni prodotte con riferimento alle istituzioni pubbliche e private.
Lo sviluppo di un Sistema informativo integrato offre la possibilità di coordinare la produzione dei dati statistici al fine di ottenere l'integrazione di tutte le informazioni disponibili all'interno della Direzione relative all'attività di ogni singola impresa (o azienda agricola). L'esistenza di un Sistema informativo integrato consente, tra l'altro, di evitare l'inutile duplicazione dei dati, con conseguente notevole ottimizzazione nel processo di raccolta ed analisi degli stessi. Infatti con l'integrazione e l'aggregazione dei dati in un unico warehouse si è riusciti ad avere un'unica fonte coerente di dati ed a distribuirli attraversol'Intranet, rendendoli fruibili attraverso un qualunque Web browser. In particolare sono stati considerati indicatori economici derivanti dalle due indagini leader delle statistiche strutturali SCI (Sistema Conti Imprese) e PMI (Piccole Medie Imprese) tra cui il fatturato, numero addetti, costi del personale e così via.
Le variabili di classificazione delle imprese sono l'attività economica, classe di addetti e regioni e a seconda del livello di dettaglio alcuni indicatori sono nascosti per gli utenti esterni per problemi di riservatezza dei dati..



Progettazione e realizzazione del Web Warehouse

L'ambiente OLTP è rappresentato sia da database Oracle sia da tabelle SAS. Una volta individuati gli indicatori economici e il livello di aggregazione abbiamo realizzato un data warehouse di primo livello a partire dal quale sono stati sviluppati data mart per consentire la distribuzione delle informazioni su Intranet.

In particolare abbiamo creato 3 data mart per tre aggregazioni diverse:

1. Divisione di attività economica per regioni;
2. Gruppo di attività economica per classe di addetti;
3. Classe di attività economica.

Per ogni classificazione sono stati stabiliti circa venti indicatori economici diversi.

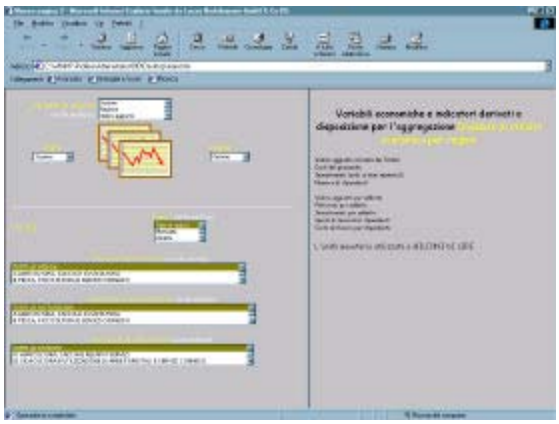
Interfaccia Web per il Warehouse

- *Accesso controllato al DW*

Sia gli utenti esterni che quelli interni devono registrarsi prima di poter visionare le informazioni del warehouse. Per ogni utente viene generata una password, memorizzata in una tabella Oracle insieme ai dati anagrafici e richiamata ad ogni accesso alla banca dati.

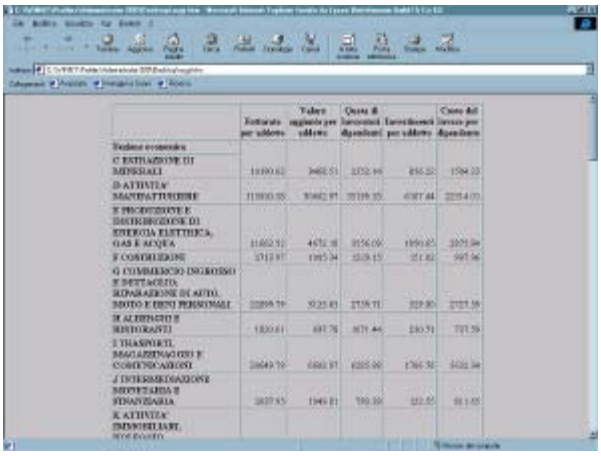
- *Scelta del tipo di informazione mediante l'uso dei metadati*

E' possibile consultare i metadati per visionare il significato delle variabili di classificazione e gli indicatori economici prima di qualsiasi richiesta.



- *Richiesta di Query e Report dinamici*

L'utente può effettuare query dinamiche in base alle variabili di classificazione disponibili e scegliere il tipo di report da scaricare definendo a priori le righe, le colonne e le pagine della tavola statistica.



- *Scarico dei report in Excel*

  E' possibile infine scaricare sul proprio PC la tavola statistica in formato Excel.

Da SAS/IntrNet v 6.12 alla nuova Versione 8.00

L'applicazione di reportistica dinamica su Web è stata realizzata con la versione 6.12 di SAS/IntrNet e aggiornata poi con la versione 8.00 per poter usufruire dei nuovi layout abbastanza flessibili di ODS (output delivery system) come tabelle 'linkate' ad altre in maniera interattiva.

Conclusioni

L'integrazione di molteplici processi produttivi tra loro molto distanti (per tecniche di rilevazione, obiettivi, codifiche, strumenti informatici, etc.) è un lavoro lento e costoso che non può essere risolto in tempi brevi, ma richiede necessariamente un'implementazione per fasi successive. Per questo l'Istat in collaborazione con SAS Institute in una prima fase ha realizzato un processo end-to-end per la realizzazione di un Data Webhouse con dati derivanti da due particolari indagini statistiche d'impresa: da un lato il sistema contiene una banca dati integrata e storicizzata, dall'altro applicativi per la navigazione multidimensionale, selezione, analisi e pubblicazione di tavole statistiche su web.

Banca Popolare di Milano (relazione presentata in occasione del Convegno SUGItalia 2000)

Abstract

Il progetto 'DRIVER' è nato per soddisfare l'esigenza dell'Istituto di monitorare l'efficienza della propria rete dopo averla riorganizzata in un concetto di CRM denominato 'Modello PRO - Professionalità, Relazioni, Obiettivi'.
Nell' ambito dei servizi operativi di agenzia vengono valutati i carichi di lavoro svolto e la misurazione temporale degli stessi, per la parte commerciale vengono analizzati, avvalendosi della metodologia della regressione statistica, numerosi indicatori di natura commerciale (raccolta, impieghi, margini, ecc.).
Il sistema è in grado di distinguere la crescente attività svolta dai nuovi canali di commercio Call-center, Inlineaweb, We@Bank e Home Banking.

Per riuscire a garantire :
- uno sviluppo in tempi rapidi di una applicazione ad hoc
- una completa ed efficace gestione dei processi che rilevano i carichi di lavoro
- un costante monitoraggio delle evoluzioni del mercato (operatività diversa, volumi...)
- la possibilità di implementare velocemente e ordinatamente il modello
- la visualizzazione dei dati raccolti in modo veloce e personalizzato all'utente

…ci si è avvalsi di prodotti SAS.
Per rilevare le operazioni svolte in dipendenza, con il relativo tempo impiegato, con la collaborazione di UP Tecno, si sono creati dei moduli 'custom' in linguaggio SAS costituendo in ambiente mainframe delle elaborazioni notturne che mensilmente estraggono i dati dai singoli partitari utilizzati dalla nostra banca.
Con il supporto di DWH Administrator si sono ordinati i flussi ottenuti dopo averli integrati ed arricchiti di alcune informazioni anagrafiche e qualitative attraverso processi direttamente disegnati e generati dal prodotto, si è creato un datawarehouse. Il datawarehouse costituito è stato messo a disposizione dell'utente che riesce ad estrarre i report da esso stesso costruiti e manutenuti in completa autonomia a diversi livelli di aggregazione (singola dipendenza, area, istituto).

Nuovo modello commerciale -> adeguamento modello di dimensionamento

A partire dal 1998 la BPM ha attivato e progressivamente esteso a tutta la rete commerciale un nuovo modello organizzativo (P.R.O. - Professionalità Relazioni Obiettivi), basato sui seguenti elementi:

- segmentazione della clientela in base a criteri dimensionali quali fatturato, patrimonio...

- ❖ grandi imprese
- ❖ corporate
- ❖ retail aziende
- ❖ retail privati plus
- ❖ retail privati standard

- differenziazione dei canali distributivi:

- ❖ grandi imprese: gestite centralmente nell'ambito della Dir. Marketing e Commerciale
- ❖ corporate: creazione di nuove unità operative Corporate
- ❖ retail: gestite dalle "vecchie" unità operative (agenzie)

- nell'ambito dei diversi segmenti,**definizione di portafogli clienti** assegnati a figure di **gestori**.

Tra le varie attività intraprese per rendere operativo il nuovo modello si è resa necessaria una **revisione del modello di dimensionamento delle agenzie.**

Caratteristiche del **"vecchio modello"** di dimensionamento (Primo Approdo):

calcola il **dimensionamento complessivo** dell'unità operativa (agenzia), senza distinguere tra attività commerciali (relative alla gestione del rapporto con il cliente) ed operative/amministrative;

non opera distinzioni tra attività riferite ai **diversi segmenti** di clientela nè ai canali distributivi;

gestisce dati operativi ad un **livello molto aggregato** - circa una ventina di macroprodotti che ricomprendono l'attività complessiva dell'agenzia (fidi, conti correnti, titoli,...)

Caratteristiche del **"nuovo modello"** di dimensionamento (DRIVER):

articolazione in **due "sub-modelli" distinti** per i comparti commerciale ed operativo;

per il comparto operativo, **articolazione su un numero elevato di indicatori** desunti direttamente dall'analisi dei principali processi di agenzia, con indicazione del segmento di clientela e del canale;

per il comparto commerciale, calcolo del dimensionamento teorico **per figura professionale** (gestori aziende/plus/standard), attraverso analisi di correlazione statistica tra variabili (numero di clienti gestiti, dati di redditività, di operatività, di mercato)

Logiche di dimensionamento del comparto operativo

Per quanto riguarda il comparto operativo, il modello calcola il dimensionamento attraverso **l'applicazione di tempi standard ad una serie di indicatori** (driver), raggruppati per categorie, che misurano lo svolgimento dei principali processi operativi di agenzia.

Esempio di raggruppamenti/driver:
ASSEGNI BANCARI
numero libretti assegni rilasciati
numero a/b utilizzati per prelievo contanti
numero a/b inviati inviati al protesto
BONIFICI
numero beneficiari bonifici evasi localmente
numero ordinanti con supporto magnetico

L'identificazione dei driverha richiesto le seguenti attività:

mappatura generale dei processi di agenzia

identificazione per ciascun processo degli indicatori principali

verifica dell'esistenza/reperibilità degli indicatori

I tempi standard sono stati stabiliti attraverso la **rilevazione "sul campo"** su un campione di agenzie e una **successiva fase di confronto** dei risultati ottenuti con i responsabili dei servizi operativi delle agenzie considerate.

Il modello applica inoltre, per il calcolo delle risorse complessive del comparto operativo, un "coefficiente di maggiorazione" espressivo dell'insieme di attività di agenzia non direttamente misurabili tramite indicatori (es. attività di gestione delle risorse, ...)

Logiche di dimensionamento del comparto commerciale

Per quanto riguarda il comparto commerciale, il modello calcola le risorse teoriche**distintamente per figura professionale** (gestori aziende, gestori privati plus e gestori standard).

Il modello può essere impostato in alternativa per calcolare le risorse

"a somma zero", nell'ottica cioè di **un'ottimale redistribuzione** tra le agenzie, lasciando invariato il numero totale di gestori;
in vista di un **recupero complessivo di efficienza**, identificando cioè una **situazione "obiettivo"** (un particolare portafoglio posizionato "sopra la media" per numero di clienti/redditività/ ...) e valutando i restanti portafogli in base ai parametri definiti da questa situazione-tipo.

Le informazioni acquisite dal modello riguardano:
il **numero di clienti gestiti** per ogni segmento/portafoglio, considerando soltanto quei clienti che rappresentano effettivamente un onere gestionale per la banca
dati economico-patrimoniali quali ad es. i volumi di raccolta e impieghi, il margine di intermediazione e di contribuzione, ecc.
alcuni **dati operativi** (fidi e prestiti personali, operazioni in titoli, operazioni estero ecc.)
dati di mercato (sportelli bancari operanti nella zona di competenza dell'agenzia, trend demografico, aziende operanti sul territorio ecc.)

Sull'insieme di tali variabili viene effettuata una distinzione tra
variabili "determinanti", che entrano cioè nell'algoritmo di calcolo, ed il relativo peso;
variabili "di corredo", che vengono comunque evidenziate nella reportistica per una lettura più completa dei dati prodotti.

Partendo dalle variabili del primo tipo il modello arriva così a definire, attraverso l' analisi di correlazione statistica, il dimensionamento teorico dei portafogli e delle unità operative.

La struttura organizzativa/tecnica del progetto

Dopo un' accurata analisi delle fonti informative della banca e un censimento dei dati da estrarre, il progetto ha preso inizio con l' attività di sviluppo dei moduli SAS 'custom' seguendo una precisa struttura in modo da poterli facilmente integrare.
I moduli sono stati strutturati tenendo in considerazione i seguenti principi :
- la semplicità
- la riusabilità
- l 'integrazione

Seguendo queste linee guida ci si è garantita la facilità di implementazione e manutenzione degli stessi, la comprensibilità e l'indipendenza in modo che se qualche partitario alimentante dovesse in futuro subire cambiamenti si possa circoscrivere l'intervento di 'correzione'.
Sono circa 40 i moduli che vanno a reperire informazioni dal sistema informativo del nostro Istituto.
Tali estrazioni fanno parte di una elaborazione notturna che viene eseguita mensilmente. I dati ottenuti vengono depositati in data set SAS per il successivo utilizzo degli stessi attraverso il codice generato da DWH.
In DWH è stato possibile sviluppare, per il resto del progetto, tutto il rimanente software necessario al fine di ottenere una tabella dettagliata per singolo cliente dove vieneripilogata l'attività del cliente a livello mensile, in ciascuna dipendenza e per tipologia di operazione, abbinata al tempo di esecuzione ottenuto applicando le regole dei criteri (area di appartenenza, dimensione, indicatori di mercato).
Naturalmente la struttura costruita permette l'inserimento di nuovi 'driver' e la modifica delle regole di abbinamento del tempo di esecuzione per ciascun indicatore.
I prossimi sviluppi del progetto sono di integrare la parte di rilevazione 'commerciale' nella struttura esistente in modo di avere un disegno completo dell' applicazione.

Costi di realizzazione del progetto

Al progetto hanno preso parte tre persone di profilo tecnico di cui due occupate a tempo parziale e alcune risorse organizzative per l'attività di macro analisi iniziale.
Avvalendosi comunque di strumenti SAS si è portato a termine il progetto nei tempi pianificati riuscendo anche a ritagliare il tempo per documentare in maniera soddisfacente l'applicazione ed affiancare per qualche giorno l'utente nella costruzione dei primi report desiderati.

WORKSHOP

Data warehousing in ambiente statistico

Considerazioni su utilizzo di tecnologie OLAP/Data Warehousing in ambiente statistico

G. D'angiolini

Considerazioni sull'utilizzo di tecnologie OLAP/Data Warehousing in ambiente statistico

Giovanna D'Angiolini

Introduzione.

L'attuale offerta di tecnologie per il Data Warehousing risponde al diffondersi dell'esigenza di utilizzare l'informazione prodotta all'interno di un'organizzazione in funzione di supporto alle decisioni. A partire dal noto articolo di Codd (Codd, 1990) con il termine OLAP (On Line Analytical Processing) si indica tale utilizzo analitico dell'informazione, denotando come OLTP (On Line Transaction Processing) le tradizionali attività di utilizzo dell'informazione in funzione direttamente produttiva. Un warehouse è in effetti un database organizzato in modo da ottimizzare l'uso dell'informazione da parte di applicazioni OLAP. Con il termine OLAP/Data Warehousing si etichetta ormai ogni strumento o soluzione architetturale mirato a rendere possibile l'utilizzo dell'informazione disponibile presso un'organizzazione per l'analisi di fenomeni inerenti la sua attività. L'analisi statistica ha un ruolo preminente nell'utilizzo analitico dell'informazione. In generale quindi il diffondersi dell' approccio OLAP/Data Warehousing determina un'offerta di tecnologie informatiche assai più adeguate agli usi statistici, rispetto alle tradizionali tecnologie database. Questo intervento è dedicato a discutere *quanto* le tecnologie OLAP/Data Warehousing siano effettivamente appropriate ad un contesto di produzione e uso di dati statistici, con particolare riferimento a due aspetti: il modello di rappresentazione dei dati e di interazione con gli utenti adottato dagli strumenti OLAP, il modello di architettura dei sistemi informativi incentrato sul concetto di warehouse.

L'approccio OLAP/Data Warehousing e i Sistemi Informativi Statistici.

Nell'approccio OLAP/Data Warehousing si assume che una specifica organizzazione, spesso un'impresa, costituisca l'universo di riferimento del warehouse e delle applicazioni OLAP che lo utilizzano. Gli utenti del warehouse assumono decisioni riguardanti l'organizzazione a diversi livelli, hanno perciò bisogno di valutare le performance dell'organizzazione in relazione al loro specifico settore o ambito decisionale, e per questo utilizzano strumenti OLAP che accedono al warehouse, nel quale sono resi disponibili per i diversi usi dati provenienti soprattutto dal sistema informativo dell'organizzazione. Il limite di questa visione è nella mancata distinzione tra i diversi livelli decisionali e le correlate diverse caratterizzazioni qualitative delle esigenze informative dei

decisori. Le modalità d'uso dell'informazione alle quali fa prevalentemente riferimento l'approccio OLAP/Data Warehousing sono proprie dei processi decisionali implicati nella gestione corrente delle attività di un'organizzazione. Il punto è che la decisione ad alto livello, orientata alla definizione di strategie, da parte di un'organizzazione o di qualsiasi altro soggetto, richiede flussi informativi più ricchi e diversamente organizzati, dato che ogni scelta strategica implica lo studio ad ampio raggio e la valutazione dello stato di quella parte del mondo reale che influenza le scelte e ne è influenzato. A supporto della decisione strategica si rende allora indispensabile un'attività parallela di osservazione e analisi dell'ambito dei fenomeni rilevante per le scelte, organizzata e condotta con continuità, indipendentemente rispetto al momento della decisione. I Sistemi Informativi Statistici sono costituiti in funzione di questa attività. Anche se l'esigenza di tenere sotto osservazione uno specifico ambito di fenomeni è comunque in ultima analisi originata dalla necessità di garantire un adeguato input informativo a specifici processi decisionali, nella progettazione di un Sistema Informativo Statistico l'obiettivo diretto è ottimizzare l'osservazione e lo studio del fenomeno in quanto tale. L'universo di riferimento non è un'organizzazione, ma direttamente il mondo reale ed uno specifico ambito di fenomeni in esso, l'utilizzatore non è direttamente un decisore ma un analista del fenomeno. Ciò è vero per tutti i Sistemi Informativi Statistici, anche quelli appartenenti a singole organizzazioni, ma si rende particolarmente evidente nel caso dei Sistemi Informativi Statistici pubblici, costituiti in funzione di supporto alla valutazione delle politiche pubbliche e/o per l'uso da parte di una vasto e spesso indefinito insieme di utenti. Ci si deve chiedere se queste differenze di finalizzazione determinino limiti all'applicabilità dell'approccio OLAP/Data Warehousing, e delle tecnologie basate su di esso, al dominio dei Sistemi Informativi Statistici. A nostro parere, un limite generale risiede nelle modalità di interazione con l'utente e nel modello di rappresentazione dei dati dei quali si avvale la maggior parte degli strumenti OLAP. Un limite più specifico riguarda più in particolare le organizzazioni produttrici di Sistemi Informativi Statistici pubblici, per le quali dev'essere attentamente discussa l'applicabilità di architetture del sistema informativo basate su warehouse.

Caratteristiche dei Sistemi Informativi Statistici.

Un Sistema Informativo Statistico è in generale costituito di più fonti d'informazione, indagini o archivi amministrativi. I Sistemi Informativi Statistici sono inoltre frequentemente estesi, aggiungendo nuovi dati o nuove fonti o integrando sistemi originariamente distinti. Ogni fonte ha associato un proprio procedimento per la rilevazione dell'informazione, ed un proprio contenuto informativo descrivibile in termini di aspetti del mondo reale osservati. Le fonti possono presentare

il massimo grado di eterogeneità, relativamente tanto alle procedure d'osservazione quanto alla descrizione del contenuto informativo, la specifica degli aspetti del mondo reale osservati da fonti diverse può basarsi su diverse concettualizzazioni. La costituzione di un Sistema Informativo Statistico presuppone l'integrazione concettuale delle fonti, vale a dire la definizione di una descrizione integrata del loro contenuto informativo, inoltre comporta la documentazione dei contenuti e delle caratteristiche delle fonti e del sistema nel suo complesso attraverso la definizione di opportune classi di metadati, e può comportare una limitata integrazione fisica di dati provenienti da fonti diverse relativi agli stessi aspetti del mondo reale osservati. L'atteggiamento dell'utenza dei Sistemi Informativi Statistici è esplorativo: l'utente individua e manipola i propri dati d'interesse, scegliendo tra le fonti e i dati disponibili sulla base degli aspetti del mondo reale cui è interessato. L'integrazione concettuale delle fonti e la disponibilità di documentazione, sotto forma di metadati, serve proprio a guidare l'utente in questa attività.

L'interazione con l'utenza e il modello di rappresentazione dei dati: le esigenze dell'utenza statistica.

Come è noto, gli strumenti OLAP consentono all'utente la manipolazione on-line dei dati estratti dal warehouse. A questo scopo pongono a disposizione dell'utente una vista dei dati gestiti nel warehouse (un data-mart), determinata sulla base delle sue specifiche esigenze informative, sulla quale l'utente può effettuare una serie di operazioni, in modo da ricavare i propri dati d'interesse, articolandone inoltre come meglio crede la presentazione. Le operazioni elementari sui dati effettuabili con gli strumenti OLAP consistono in aggregazioni e disaggregazioni (roll-up e drill-down), selezione di specifici elementi (slicing). E' importante sottolineare come l'esigenza di elaborare on-line l'informazione disponibile si ponga negli stessi termini per l'utenza dei Sistemi Informativi Statistici: il diffondersi di strumenti OLAP porta quindi alla luce un bisogno dell'utilizzatore di dati statistici mentre offre la tecnologia per soddisfarlo. Rispetto ai bisogni dell'utenza di Sistemi Informativi Statistici, i limiti di questi strumenti non risiedono in ciò che permettono di fare, ma nel come. Nel caso dei Sistemi Informativi Statistici, le esigenze informative dell'utenza non sono delimitabili a priori, oltre a ciò mutano frequentemente le figure utente ed il contenuto informativo stesso del sistema. Di conseguenza, al posto dell'accesso ad uno specifico data-mart, dev'essere resa possibile all'utente un'articolata attività di esplorazione del sistema, tesa a costruire on-line il proprio data-mart. Nei Sistemi Informativi Statistici quindi l'interazione tra utente e sistema è in primo luogo finalizzata a guidare la navigazione e l'identificazione dei dati d'interesse, in secondo luogo a permettere la manipolazione on-line dei dati come nei sistemi OLAP. La navigazione si svolge mediante un articolato colloquio utente-sistema nel quale sono

richiesti e trasmessi metadati descrittivi del contenuto informativo disponibile. In questo colloquio, l'articolazione in classi e l'organizzazione dei metadati, così come il linguaggio utilizzato, hanno un ruolo fondamentale. Entrambi si basano necessariamente su un modello concettuale per la rappresentazione e la descrizione dei dati disponibili, che guidi la specifica di tutti gli aspetti rilevanti nella definizione del dato. L'efficacia dell'interazione dipende dall'adeguatezza della concettualizzazione alla base del modello di rappresentazione dei dati utilizzato, che deve consentire di caratterizzare completamente i dati in funzione del loro utilizzo statistico, permettendone quindi anche la comparazione e l'uso congiunto. Per lo statistico, si è detto, i dati devono essere descritti in termini di aspetti del mondo reale osservato. Lo statistico assegna i diversi aspetti del mondo reale a diverse categorie concettuali in funzione del loro ruolo nell'analisi, distinguendo unità d'analisi, variabili, classificazioni. Un modello di rappresentazione dei dati adeguato all'uso statistico dell'informazione è quindi un modello che utilizza queste categorie. L'uso di categorie concettuali familiari allo statistico non è però sufficiente; nel contesto di Sistemi Informativi Statistici composti di una pluralità di fonti eterogenee, i dati sono anche caratterizzati dal complesso delle relazioni che li legano: relazioni di derivabilità tra dati, variabili, classificazioni, relazioni concettuali tra dati dipendenti da relazioni esistenti tra gli aspetti del mondo reale osservati.
Si è detto che nei Sistemi Informativi Statistici la comparabilità tra dati eterogenei è assicurata dall'integrazione concettuale tra le fonti: la descrizione del contenuto informativo oggetto della comunicazione tra utente e sistema è infatti una descrizione integrata, formulata in termini di concetti omogenei. Per assolvere al proprio ruolo, tale descrizione dev'essere ricavata utilizzando un modello di rappresentazione di adeguata espressività, nel senso appena precisato. La definizione di un appropriato modello di descrizione e rappresentazione dei dati, specifico per i Sistemi Informativi Statistici, è quidi lo snodo attorno al quale si articolano funzioni fondamentali: l'attività di integrazione concettuale, la definizione delle classi rilevanti di metadati e la loro organizzazione, la comunicazione utente-sistema finalizzata all'individuazione e alla manipolazione dei dati d'interesse. Sulla definizione di questo modello il dibattito è aperto: il punto è che esso sembra in ogni caso dover essere più ricco del modello di rappresentazione dei dati sul quale si basano gli strumenti OLAP, anche se al prezzo di una minore intuitività.

L'interazione con l'utenza e il modello di rappresentazione dei dati: i limiti degli strumenti OLAP.

Come è noto, nei sistemi OLAP il modello utilizzato per la descrizione dei dati disponibili si basa sulla metafora del cubo multidimensionale. Un dato è descritto come un aggregato di misure e

dimensioni. Le misure rappresentano quantità osservate e costituiscono l'oggetto della manipolazione. Le dimensioni costituiscono caratteristiche qualitative alle quali sono associate misure. Alle dimensioni possono essere associate gerarchie di livelli. Ad esempio, il dato "Numero e fatturato delle imprese per attività economica e forma giuridica" è descritto da un cubo avente come misure il Fatturato e e il Conteggio del numero di imprese, come dimensioni la Forma Giuridica e l'Attività economica, a quest'ultima può essere associata una gerarchia di livelli di aggregazione, ad esempio quella offerta dalla classificazione ATECO 91. Il cubo multidimensionale è una rappresentazione del dato molto funzionale alla manipolazione: ad esempio, l'aggregazione (roll-up) e la disaggregazione (drill down) sono rappresentate come eliminazione e aggiunta di dimensioni, rispettivamente. Il modello del cubo multidimensionale ha il merito di non confondere la definizione del dato, visto come oggetto manipolabile, con la sua presentazione in due dimensioni, come insieme di tavole. Non è però adeguato a convogliare all'utente statistico una rappresentazione del dato conforme alle sue esigenze, e cioè espressa in termini degli aspetti del mondo reale osservati ai quali il dato è riferito. Ciò limita non solo la possibilità di confrontare e scegliere tra dati, ma anche la possibilità di offrire supporto a funzioni di manipolazione particolarmente interessanti per lo statistico, che coinvolgono dati differenti: tra queste, il calcolo di indicatori, che in termini OLAP equivale a porre in relazione due cubi, definendo una nuova misura come rapporto tra misure estratte dai due cubi. Questa inadeguatezza non è casuale: nei sistemi OLAP non si assume che l'utente abbia l'esigenza di operare una scelta tra fonti e tra dati differenti, cade quindi la necessità di una descrizione dei dati completa dal punto di vista del riferimento agli aspetti del mondo reale osservati: la descrizione dei dati è unicamente finalizzata alla loro manipolazione. E' implicito il riferimento dei dati ad una realtà delimitata, quale un'organizzazione, un contesto nel quale è realistico assumere che l'utente di un data-mart possieda già una conoscenza adeguata del significato dei dati che abitualmente utilizza.

Alcuni esempi sono utili per evidenziare i limiti del modello di rappresentazione adottato dagli strumenti OLAP.

- Due dati aggregati possono essere composti delle stesse misure e dimensioni, ma essere riferiti a unità d'analisi differenti (come per "numero di iscitti all'università per sesso e anno di corso", "numero di iscritti all'università stranieri per sesso e anno di corso"). Come nelle più comuni rappresentazioni tabellari, nelle quali il nome dell'unità d'analisi compare nel titolo della tabella, l'unità d'analisi può comparire nel nome del cubo, è evidente d'altra parte che questa soluzione non garantisce quella rappresentazione esplicita e standardizzata di tutte le componenti rilevanti del dato che è indispensabile all'utente per individuare la propria informazione d'interesse.

- Il concetto OLAP di dimensione non distingue tra variabile e classificazione, cioè tra una proprietà dell'unità d'analisi che interviene nella definizione del dato, e l'insieme dei suoi possibili stati. Si considerino ad esempio i due dati "numero degli iscritti all'università per sesso e regione di localizzazione della sede universitaria", "popolazione in età 19-25 anni per sesso e regione di residenza". In questo esempio, i due dati condividono la variabile sesso con la relativa classificazione, mentre utilizzano la stessa classificazione, data dall'insieme delle regioni italiane, per due variabili diversamente definite, "regione di localizzazione della sede universitaria" e "regione di residenza". Sulla necessità di distinguere tra variabili e classificazioni valgono naturalmente considerazioni analoghe a quella precedente, relativa all'unità d'analisi. Considerazioni ulteriori riguardanti il calcolo degli indicatori confermano questa necessità. Dai due dati precedenti è ottenibile l'indicatore "iscritti all'università per 100 giovani in età 19-25, per sesso e regione". La descrizione di questo dato in termini di aspetti del mondo reale di riferimento non è tutta espressa nel nome, dipende invece dalla definizione dei dati da cui è derivato, e nella sua definizione è importante la specifica delle due diverse proprietà alle quali è riferita la classificazione per regioni, rispettivamente per il numeratore e il denominatore del rapporto. Molti sistemi OLAP non consentono la navigazione tra cubi differenti, e quindi la costruzione on-line di indicatori. Anche per quelli che la consentono d'altra parte l'uso del concetto di dimensione comporta limitazioni o alla manipolabilità, o ad una corretta espressione del significato dei dati ottenuti. Nel caso precedente, per poter ottenere l'indicatore è necessario definire "regione" come dimensione in comune: in questo modo però non è correttamente espresso il significato dei dati di partenza, che dipende dalla coppia variabile-classificazione, nè di conseguenza quello del dato ottenuto.
- Si consideri il dato "numero degli iscritti all'università per sesso e tipologia del corso di laurea". Nella definizione di questo dato è evidente come il significato delle dimensioni sia differentemente costruito: la variabile "sesso" appartiene direttamente all'unità d'analisi "iscritti all'università", mentre la variabile " tipologia del corso di laurea " appartiene all'unità di analisi "corso di laurea", ed è attribuibile all'unità d'analisi "iscritti all'università" solo attraverso il legame concettuale (l'iscrizione) che intercorre tra le due unità. Si consideri un'indagine nella quale si siano raccolte informazioni relative ai corsi di laurea, e, per ciascuno di essi, relative agli iscritti. In questo caso al legame concettuale tra i corsi di laurea e i loro iscritti corrisponde una relazione concreta tra i relativi dati osservati. In questa indagine, il dato "numero degli iscritti all'università per sesso e tipologia del corso di laurea " è ottenibile navigando lungo la relazione che lega il dato osservato relativo agli studenti a quello relativo al corso di laurea. Relazioni di questo tipo non sono rappresentate nei sistemi OLAP, e non è quindi possibile all'utente navigare

on-line lungo i legami tra dati disponibili per costruire un nuovo dato d'interesse. I dati ottenibili per navigazione da dati osservati concettualmente connessi vengono precostituiti al momento della creazione dei data-mart e direttamente presentati all'utente. Nei Sistemi Informativi Statistici, data la non predeterminabilità delle esigenze dell'utenza, questo comporterebbe la predeterminazione di tutti i dati ottenibili per navigazione tra dati connessi. Ciò non solo è operativamente complesso quando i dati sono collegati tra loro da relazioni numerose e ramificate, ma soprattutto rende più difficile trasmettere all'utente la ricostruzione del significato del dato in termini di aspetti del mondo reale osservato. Da questo punto di vista, permettere all'utente di costruirsi on-line un nuovo dato per navigazione tra dati connessi ha il vantaggio di consentire anche una costruzione della semantica del dato ottenuto articolata per passi successivi. Questo modo di procedere richiede un modello di rappresentazione dei dati che permetta la rappresentazione esplicita dei legami tra dati osservati. Si perde in questo modo l'immediatezza di accesso al dato e di manipolazione garantita dai sistemi OLAP, che d'altra parte si perderebbe comunque, nei Sistemi Informativi Statistici, a causa della complessità del mondo reale sottostante i dati manipolati.

Come in molti altri casi, la crescità in quantità e complessità dell'informazione pone problemi di adeguamento non solo quantitativi, ma anche qualitativi degli strumenti di supporto all'utilizzo. L'analisi precedente è volutamente focalizzata sul caso più complesso, nel quale il Sistema Informativo Statistico è composto di numerose fonti eterogenee, e le esigenze dell'utenza non sono predeterminabili. Un'analisi più attenta dell'utenza dei Sistemi Informativi Statistici porta in effetti ad individuare diverse tipologie di utenza, per alcune delle quali l'utilizzo degli strumenti OLAP offerti sul mercato è perfettamente adeguato. Occorre osservare, d'altra parte, che proprio i Sistemi Informativi Statistici pubblici tendono oggi a presentare con sempre maggiore frequenza le caratteristiche di complessità descritte.

Warehouse e SIS pubblici

Nell'approccio OLAP/Data Warehousing, un unico data warehouse è costituito per l'organizzazione nel suo complesso, dal quale sono definiti i data-mart destinati all'utilizzo da parte di diverse classi di utenza. A differenza dei database costituiti per l'uso da parte di applicazioni OLTP, un warehouse gestisce serie storiche di dati ed è aggiornato a specifiche cadenze temporali. Se si usa il termine warehouse per denotare genericamente una raccolta di dati così caratterizzata, i Sistemi Informativi Statistici possono descriversi come sistemi per l'accesso a warehouse di dati relativi a specifici ambiti di fenomeni.

Compito delle organizzazioni della statistica ufficiale è oggi in misura crescente la produzione di una pluralità di Sistemi Informativi Statistici pubblici relativi a diversi ambiti di fenomeni. Gli ambiti di fenomeni posti sotto osservazione sono funzione del soddisfacimento delle esigenze informative che via via si manifestano. I confini dei Sistemi Informativi Statistici pubblici possono perciò essere determinati in base a diversi criteri: potranno essere realizzati sistemi di tutti i dati provenienti dalle indagini congiunturali sulle imprese o di tutte le fonti di dati sulla sanità, così come sistemi specificamente dedicati all'osservazione della condizione degli anziani o del mercato del lavoro.

Per i Sistemi Informativi Statistici pubblici quindi il rapporto con l'organizzazione è rovesciato, rispetto al caso delle applicazioni OLAP. Un Sistema Informativo Statistico pubblico non è relativo ai fenomeni che interessano una singola organizzazione nè utilizzato esclusivamente in essa, è invece l'organizzazione a svolgere attività di raccolta dell'informazione destinata ad alimentare sistemi informativi di utilizzo generale, attraverso l'organizzazione di indagini o l'acquisizione di archivi amministrativi. Questo determina un primo punto di differenziazione della realtà dei Sistemi Informativi Statistici pubblici rispetto agli scenari presi in considerazione dall'approccio OLAP/Data Warehousing, che riguarda in particolare le modalità di costituzione e alimentazione del warehouse. L'aggiornamento del warehouse di un'organizzazione avviene estraendo i dati da database e file costituiti in funzione di applicazioni OLTP, o altrimenti acquisiti, e opportunamente trasformandoli. Nel caso dei Sistemi Informativi Statistici pubblici, i dati gestiti provengono da indagini o archivi amministrativi acquisiti e modificati in vista del loro utilizzo statistico. Alcune soluzioni tecniche concepite per la costituzione e l'alimentazione di warehouse possono essere sfruttate per i Sistemi Informativi Statistici pubblici; d'altra parte, è evidente che il problema specifico che si pone per questi ultimi è assolutamente peculiare, e riguarda l'articolazione tra i sistemi di supporto alla produzione statistica, in particolare di supporto all'esecuzione di indagini, e l'insieme dei Sistemi Informativi Statistici pubblici prodotti.

Un secondo, più importante punto di differenziazione tra applicazioni OLAP/Data Warehousing e Sistemi Informativi Statistici pubblici si ravvisa nella diversità dell'ambito e degli scopi dell'integrazione. Il warehouse di un'organizzazione è unico e integrato, sia da un punto di vista concettuale che da un punto di vista fisico: non esistono duplicazioni di dati riferiti agli stessi aspetti dell'organizzazione. Nelle applicazioni statistiche, i dati provenienti da indagini diverse non vengono integrati fisicamente, ma eventualmente sottoposti a procedure di linkage, le quali comportano una specifica caratterizzazione dell'informazione ottenuta. E' inoltre irrealistico pensare che un'organizzazione della statistica ufficiale possa in ogni momento garantire l'integrazione concettuale del contenuto informativo di tutti i Sistemi Informativi Statistici prodotti, anche perchè

l'attività di produzione di descrizioni integrate dei contenuti informativi di tali sistemi è più complessa di quanto non appaia, per la necessità di sviluppare un'analisi in profondità dei concetti implicati. D'altra parte, le organizzazioni della statistica ufficiale non possono ignorare l'importanza di produrre comunque una documentazione unificata dell'informazione complessivamente offerta. Si tratta di rispondere ad un'esigenza di orientamento espressa dall'utenza finale dei Sistemi Informativi Statistici pubblici, ma anche di offrire supporto alle decisioni dei progettisti di nuove indagini o nuovi Sistemi Informativi Statistici, ed a quelle degli organismi incaricati di definire le strategie di produzione dell'informazione offerta al pubblico. A questo scopo è indispensabile la costituzione di un sistema unico di documentazione dei contenuti informativi delle indagini e dei Sistemi Informativi Statistici, nel quale sia documentato anche lo stato di relativa integrazione delle fonti disponibili, e che offra strumenti di supporto all'integrazioneconcettuale tra fonti. Il sistema di documentazione diventa allora anche la fonte dei metadati utilizzati per guidare l'accesso ai dati all'interno dei singoli Sistemi Informativi Statistici pubblici.

All'integrazione fisica e concettuale dei dati garantita dal warehouse, si sostituisce quindi nel caso dei Sistemi Informativi Statistici pubblici la rappresentazione unificata delle conoscenze sui dati garantita dal sistema di documentazione. Un importante oggetto di riflessione è oggi la definizione di un'architettura generale dei sistemi informativi per le organizzazioni della statistica ufficiale, che veda da una parte l'organizzazione dei Sistemi Informativi Statistici prodotti attorno ad un sistema unico di documentazione dei contenuti informativi, dall'altra un'articolazione ottimale tra i Sistemi Informativi Statistici prodotti e i sistemi di supporto alla produzione statistica.

Riferimenti bibliografici

Codd, E. F. (1990) Providing OLAP to user-analysts: an IT mandate, *Technical Report, Codd and Associates*

ONU-ECE (1995)- Guidelines for the modelling of Statistical Data and Metadata

Kimball, R. (1996) The data warehouse toolkit,*J. Wiley&Sons*

M. Jarke, M. Lenzerini, V. Vassiliou, P. Vassiliadis (2000), Fundamentals of Data Warehouses, *Springer*

A. Shoshani, (1996) Statistical databases and OLAP: similarities and differences,*Proc. International Conference on on Information ad Knowledge Management*

B. Sundgren, (1991) Some properties of statistical information: Pragmatic, Semantic and Syntactis", *Statistics Sweden*

WORKSHOP

Data warehousing in ambiente statistico

Il data warehouse statistico, come fonte per la disseminazione dell'informazione e controllo di qualità

E. Giovannini, A. Sorce

WORKSHOP

Data warehousing in ambiente statistico

I servizi decisionali nella Pubblica Amministrazione piemontese

G. Bonello

I servizi decisionali nella Pubblica Amministrazione piemontese

Giuliana Bonello- CSI-Piemonte

Le Pubbliche Amministrazioni, nel loro ruolo istituzionale, devono monitorare ed analizzare i diversi aspetti della società di cui sono responsabili.
In tale ottica hanno il diritto-dovere di rilevare informazioni specialistiche sui diversi fenomeni.
Nel corso degli anni alcune Pubbliche Amministrazioni hanno rilevato molte informazioni ed oggi possiedono un patrimonio informativo consistente, che si presta ad essere analizzato per fini decisionali.
La Pubblica Amministrazione individua inoltre nella tecnologia uno strumento indispensabile per accrescere l'efficienza della sua azione garantendone, nel contempo, la trasparenza. La tecnologia delle reti offre oggi strumenti dotati di notevole potenza e flessibilità, che devono essere calati in un modello capace di coniugare la loro evoluzione al processo di riforma in atto. Ciò significa che qualsiasi sistema informativo, destinato a favorire i compiti di programmazione e di pianificazione assegnati all'ente pubblico dall'attuale processo di delega, presuppone un'architettura dai requisiti molto stringenti. Un'architettura che da un lato sia in grado di supportare i processi decisionali e dall'altro garantisca l'interconnessione del sistema con gli ambienti esterni e con il centro. Queste sono le considerazioni che hanno guidato il CSI-Piemonte nella realizzazione di progetti di data warehouse rivolti alle pubbliche amministrazioni.
Progetti che mirano a valorizzare l'intero patrimonio informativo fino ad oggi accumulato, che spazia dai sistemi GIS per i dati cartografici alle soluzioni che operano su dati alfanumerici o testuali, come la banca dati delle leggi regionali.
L'obiettivo di fondo dei progetti è quello di passare dai dati alle informazioni, in altre parole di favorire la transizione dai dati operativi, generati da esigenze di tipo settoriale, alle informazioni destinate a ottimizzare i processi informativi e decisionali trasversali all'intera pubblica amministrazione.

Il data warehouse della Regione Piemonte

Alla fine del 1997 la Regione Piemonte ha avviato la fase di costruzione dello strato informativo - decisionale (Data Warehouse) del Sistema Informativo Regionale (SIRe).
L'iniziativa è collegata alla necessità della Regione di disporre di una componente di sistema informativo trasversale rispetto a tutto l'ente, atta a rispondere sia ad esigenze di supporto ai processi decisionali dell'istituzione, sia ad esigenze di tipo informativo verso soggetti esterni.
Il Data Warehouse rappresenta l'approccio in grado di mettere a fattor comune il patrimonio informativo costruito nel corso degli anni, attraverso la definizione di opportune regole di integrazione. Soddisfa inoltre l'esigenza di rispondere ad una sempre crescente richiesta di informazioni da parte degli organismi più diversi (altre amministrazioni, mun icipalizzate, associazioni, ordini professionali, ricercatori, cittadini, ecc.).
Una delle componenti realizzate è quella relativa alla navigazione sulle metainformazioni nel Data Warehouse: è l'applicazione che consente di dare immediata visibilità a tutto ciò che si sta costruendo a livello decisionale. Tutte le direzioni regionali possono contribuire a rendere più ricca la "vetrina", producendo informazioni e servizi di accesso; questo favorisce lo scambio di informazioni e consente di migliorare i processi.

Nel SIRe (Sistema Informativo Regionale) della Regione Piemonte esistevano, al 1997, già numerose componenti tipiche di un'architettura di data warehouse, anche se non ancora organicamente integrate in un quadro architetturale complessivo. Era perciò necessario più uno sforzo di integrazione che una costruzione ex-novo.
Gli obiettivi primari erano due:

- *supportare processi decisionali*: questo è l'utilizzo che dovrebbe essere preminente nel caso della Regione Piemonte, proprio per le caratteristiche ed i compiti dell'ente ; per

supporto decisionale si intende sia quello rivolto alle decisioni strategiche, sia quello più propriamente relativo ai processi amministrativi;

- *diffondere informazioni* a chi necessita, secondo i livelli di accesso consentiti (altri uffici regionali, enti esterni, ecc.); poiché le informazioni sono potenzialmente diffondibili all'esterno dell'ente Regione, è importante che nelle fasi di accesso ai dati siano definite le regole d'uso e di interpretazione/elaborazione. Inoltre in relazione alla crescente esigenza di scambio e condivisione di informazioni fra enti differenti della Pubblica Amministrazione, legata al processo di decentramento amministrativo attualmente in corso, assume notevole importanza l'esigenza di garantire e rafforzare lo scambio informativo tra Enti diversi.

Il progetto di carattere trasversale ha preso in considerazione i seguenti aspetti:

- *Aspetti metodologici*: sono stati individuati metodi di lavoro comuni per i diversi progetti decisionali che le direzioni regionali avviavano; in particolare:
  - metodologia specifica da adottare da parte dei progetti di direzione (definita in modo da consentire il rispetto degli standard di qualità richiesti dall'ISO 9001),
  - definizione del modello logico dei metadati di business
  - linee guida relative agli aspetti di disegno logico e fisico degli archivi decisionali.
- *Azioni legate all'impianto del sistema*: sono state acquisite le macchine e gli ambienti software necessari
- *Azioni organizzative*: l'impianto del sistema e la realizzazione delle componenti software è stato accompagnato da azioni formative e di sensibilizzazione rivolte a tutte le Direzioni regionali: seminari di introduzione ai concetti di Data Warehouse e di presentazione del progetto in corso sono stati svolti nell'arco di due anni e rivolte alle diverse direzioni regionali

Parallelamente sono state realizzate con le singole direzioni le basi dati decisionali del Data Warehouse e le procedure di accesso ed analisi.

Infine, in stretta collaborazione con il settore Sistemi Informativi e Informatica della Regione Piemonte, il CSI-Piemonte ha realizzato un 'catalogo' delle applicazioni esistenti: un modulo che permette agli utilizzatori di accedere a una 'vetrina' Web del Data Warehouse, di cercare le informazioni di proprio interesse tramite percorsi guidati e di attivare la relativa applicazione.
Tale applicativo di navigazione sulle informazioni e sui servizi del Data Warehouse è denominato "*Information Directory* e consente di accedere a servizi decisionali delle diverse tipologie: alfanumerica, cartografica, testuale, multimediale.
In sintesi gli obiettivi sono:

- catalogare le informazioni di interesse ed offrire uno strumento di orientamento complessivo sul patrimonio informativo della Regione;
- a tale sistema di catalogazione delle informazioni collegare i servizi di accesso ed analisi disponibili per la fruizione delle informazioni stesse;
- gestire il livello di accesso e di sicurezza delle informazioni e dei servizi offerti, attraverso profili utente differenziati, sulla base dei servizi dell'interfaccia di accesso al SIRe (Sistema Informativo Regionale).
- consentire una navigazione "personalizzabile" in base ad esigenze dell'utenza (canali di ricerca aggiuntivi, segnalibri, segnalazione di argomenti di interesse)

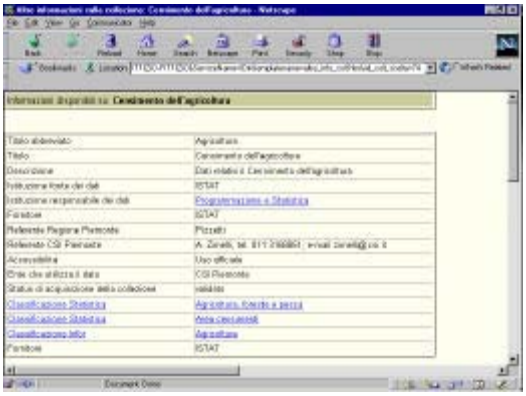
L'applicazione, classificabile come Information Portal, consente di navigare interattivamente sulle tavole dei metadati di business alla ricerca di ciò che è di interesse.
E' possibile cercare per argomento, per direzione regionale, per area territoriale e per periodo, oltre che effettuare una ricerca testuale libera.
Il risultato della ricerca sono una o più "collezioni logiche di informazioni" (equiparabili ciascuna ad un soggetto di un Data Warehouse), che comprendono archivi, prodotti finali (grafici, tabelle, pagine html, ecc.) e servizi interattivi di accesso ed analisi sui dati stessi.
Ogni "collezione logica" viene descritta nei suoi contenuti attraverso un insieme di metadati di "business".
Se la collezione logica è di interesse ed è collegata a servizi di accesso di tipo web, è possibile attivare il servizio e analizzare le informazioni.

Tra le collezioni attualmente disponibili citiamo le seguenti:

- Corsi, allievi e spesa inerente la formazione professionale
- Analisi statistiche sul lavoro (livelli di occupazione e di disoccupazione, ecc.) di interesse per l'Osservatorio mercato del lavoro
- Infortuni sul lavoro da fonte INAIL
- Osservatorio imprese artigiane
- Demografia (Banca dati demografica evolutiva sulla struttura della popolazione piemontese, archivi di fonte ISTAT)
- Finanziamenti alle aziende agricole
- Informazioni sulle scuole derivanti da questionari sulla rilevazione scolastica

Figura 1 - Catalogo di navigazione sul Data Warehouse della Regione Piemonte

Figura 2 - Accesso ai dati decisionali: Regione Piemonte "I metadati"



Il data warehouse del Comune di Torino

Nel corso degli ultimi anni la Città di Torino ha recepito le esigenze di diversi settori dell'ente di poter operare in modo più diretto sul patrimonio informativo disponibile al fine di snellire alcune procedure a supporto dei servizi e di rendere più fruibile la notevole mole di informazioni in possesso della Città, sia in modalità tradizionale (pubblicazioni) sia in modalità innovativa (via web).
All'interno del Sistema Informativo Comunale sono state realizzate componenti di carattere informativo-decisionale finalizzate a due obiettivi:

- supportare le decisioni, da un lato nell'ambito di un'ottica di controllo direzionale, dall'altro di controllo sul territorio comunale e sulle politiche relative *(logica di controllo decisionale)*
- diffondere informazioni ai diversi livelli di utenza *(logica di servizio al cittadino)*

Alcune delle componenti realizzate rendono pubbliche via Internet informazioni aggiornate di interesse generale. In dettaglio esse sono:

- Modulo di Ricerche Individuali su Dati Anagrafici
- Pubblicazione automatica mensile sul WEB della città di Torino di dati ed indici demografici desunti dall'anagrafe cittadina
- Pubblicazione automatica mensile sul WEB della città di Torino della Anticipazione degli Indici dei Prezzi al Consumo
- Procedura di Analisi dei Carichi di Lavoro nell'Accettazione delle Pratiche del Commercio
- Procedura S.I. Infrazioni per la ricerca di informazioni a carattere individuale, per l'analisi multidimensionale di dati statistici e per la produzione di report su dati delle infrazioni a partire dal 1993
- Procedura di Analisi dei Dati Elettorali

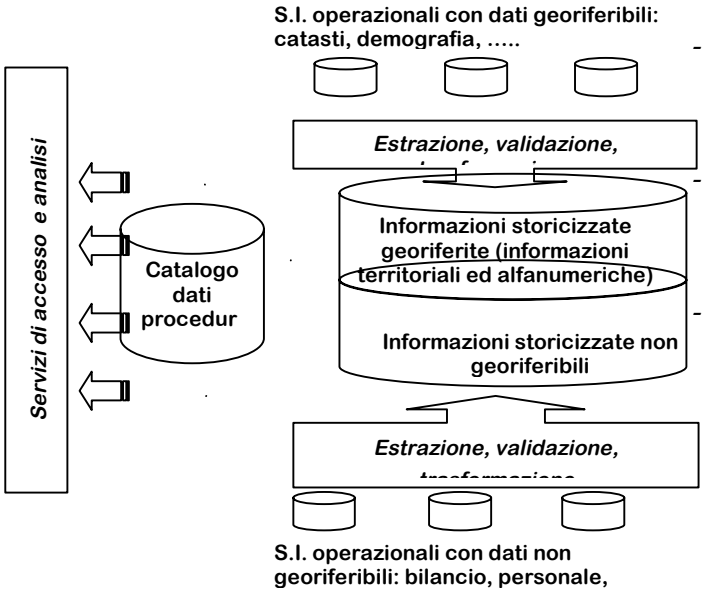
Nel corso del 2000 si è partiti con una seconda fase di progetto mirata alla realizzazione di un'infrastruttura di riferimento a livello comunale che favorisca la costruzione di un Data Warehouse organico a livello dell'intero ente, nel quale si innestino sia le applicazioni di carattere decisionale già realizzate, sia quelle in fase di sviluppo.
A regime il progetto consentirà di disporre di:

- un quadro architetturale di riferimento hardware e software specializzato per le componenti decisionali di tipo alfanumerico
- regole trasversali al SIC per l'alimentazione del Data Warehouse comunale
- un ambiente hardware e software che consenta di fruire delle nuove applicazioni decisionali con prestazioni adeguate

I progetti di interesse settoriale che verranno innestati su questa architettura sono:

- sviluppo di applicazioni a supporto dei censimenti 2001
- sviluppo di applicazioni di analisi statistica per i diversi osservatori: Osservatorio Elettorale, Osservatorio Commercio, Osservatorio Assistenza, Osservatorio Corpo di Polizia Municipale
- costruzione di banche dati decisionali contenenti informazioni socio-economiche georiferite e storicizzate, costantemente aggiornate nel tempo su cui rendere disponibile uno strumento di accesso ed analisi
- sviluppo di una base dati decisionale e di servizi di reportistica sul sistema informativo del personale
- sviluppo di strumenti per l'analisi finanziaria ed il controllo di gestione

Figura 3 – Schema logico del Data Warehouse del Comune di Torino



Altri data warehouse per la Pubblica Amministrazione Piemontese

Sono stati inoltre realizzate componenti decisionali anche per altri enti della pubblica amministrazione piemontese.

Tra queste esperienze si cita il Data Warehouse per la gestione economica del personale, realizzato come servizio di supporto per tutti gli enti che si appoggiano al CSI-Piemonte per il servizio Stipendi.
L'obiettivo è quello di fornire uno strumento flessibile e di facile utilizzo per accedere direttamente alle informazioni di carattere economico relative al personale.
Il servizio prevede l'aggiornamento mensile, necessario a mantenere efficiente il sistema informativo, della base dati e assicura un'assistenza completa: da quella formativa in fase di avviamento del Servizio, alla disponibilità quotidiana di esperti per ogni chiarimento necessario.
Il Servizio offerto prevede:

- il mantenimento dei dati storici sino a tre anni
- criteri di estrazione preimpostati
- elaborazioni "in proprio" sui dati disponibili
- integrazione con strumenti di produttività individuale

Le informazioni disponibili sono suddivise in quattro aree tematiche:

- Dati anagrafici: contiene tutte le informazioni di carattere anagrafico
- Imponibili: contiene tutte le informazioni di carattere anagrafico e le informazioni di carattere economico relative agli imponibili
- Voci: contiene tutte le informazioni di carattere anagrafico e le informazioni di carattere economico relative alle voci stipendiali
- Capitolo, lordo, ritenuta e onere: contiene tutte le informazioni di carattere anagrafico e le informazioni di carattere economico relative sia alle voci stipendiali sia agli imponibili

Figura 4 – Servizi decisionali sulla gestione economica del personale

Verso un Data Warehouse per la Pubblica Amministrazione piemontese

Il continuo decentramento di incarichi dall'Amministrazione centrale verso quella locale e l'interesse sempre più diffuso a conoscere i vari fenomeni che avvengono sul territorio regionale, stanno accelerando il processo di interscambio di informazioni tra le Pubbliche Amministrazioni, utilizzando canali come la RUPAR (Rete Unitaria della Pubblica Amministrazione Regionale).

In questa ottica il compito del CSI-Piemonte è perciò quello di realizzare sistemi informativi per la pubblica amministrazione piemontese che nascano per rispondere alle esigenze dei singoli enti, ma che consentano, ove opportuno, di essere integrati.

I livelli di integrazione necessari sono:

- all'interno della pubblica amministrazione locale
- tra la pubblica amministrazione locale e quella centrale

Inoltre l'azione concertata tra amministrazione centrale ed amministrazioni locali passa anche attraverso la creazione di legami più stretti tra la pubblica amministrazione locale e le realtà regionali di enti centrali (quali ad esempio INPS e INAIL).

Occorre perciò realizzare progetti innovativi utili a sviluppare *complessivamente* il sistema della Pubblica Amministrazione piemontese, considerando in particolare la logica dell'uso da parte degli Enti Locali.

In quest'ottica le componenti di Data Warehouse giocano un ruolo cruciale in quanto offrono la possibilità di far circolare *informazioni* tra i diversi enti, specialmente quelle realizzate in modalità web.

Il CSI-Piemonte nel 2000 ha avviato la costruzione di una *server farm per la pubblica amministrazione piemontese*, in cui trova spazio anche una componente decisionale.

Il progetto complessivo prevede un'architettura server centralizzata, sulla quale sono ospitati i database e le applicazioni, organizzata per domini; uno dei domini è quello dedicato alle applicazioni di tipo cooperativo destinate a supportare l'interscambio tra amministrazioni pubbliche.

La componente decisionale è una tipica componente cooperativa, in quanto contiene spesso informazioni raccolte da un ente ma di interesse anche per altri enti.

L'architettura identificata per la componente decisionale è a n livelli logici (vedasi figura 6), articolata in:

- database server (distinti in operational server, operational data store server, data warehouse server)
- application server (distinti in application server decisionale, web server)
- stazioni di lavoro (distinte in thin e fat client)

Per quanto riguarda l'architettura di dettaglio delle componenti decisionali sono individuabili le seguenti sottocomponenti:

- *Alimentazione del DB decisionale :* questi flussi di mapping e sommarizzazione vengono realizzati in parte con programmi ad hoc (SAS o PL/SQL) ed in parte con un tool di ETL (SAS\Warehouse Administrator). Le basi dati vengono memorizzate in Oracle con struttura R-OLAP e/o in SAS con struttura M-OLAP.
- *Gestione dei metadati*: è stato identificato un modello generale dei metadati di business del Data Warehouse;
- *Fruizione della base dati decisionale*: sono distinguibili le seguenti categorie di applicativi di fruizione dei dati :

- Query e reporting
- Analisi multidimensionale
- Applicativi specializzati di tipo statistico, analisi what-if, analisi di scenario

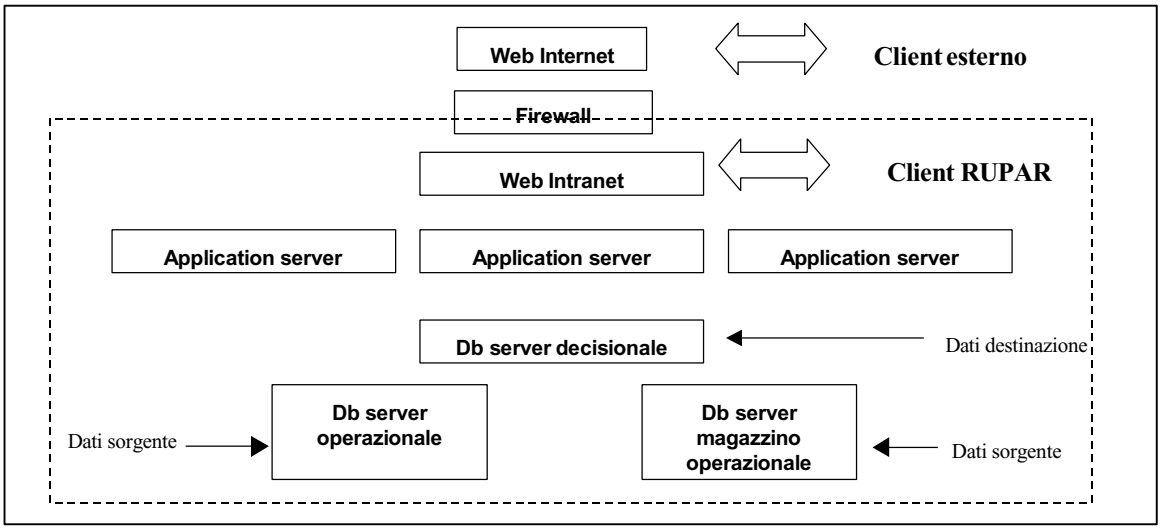
Le esigenze più orientate alla costruzione dinamica di report prevedono l'utilizzo di Business Objects e Web Intelligence, mentre le esigenze più complesse vengono soddisfatte attraverso SAS.

Conclusioni e prospettive

Uno degli obiettivi che la Regione Piemonte si pone è la diffusione di servizi sulla Rete Unitaria della Pubblica Amministrazione Regionale (RUPAR); a tale scopo la Regione Piemonte ha da tempo siglato una convenzione con l'Autorità per l'Informatica nella Pubblica Amministrazione (AIPA) al fine, tra l'altro, di sviluppare forme di collaborazione connesse alla progettazione e sperimentazione della RUPA a livello regionale e l'interconnessione della stessa con la Rete Unitaria delle Pubbliche Amministrazioni .

Per promuovere innovative forme di collaborazione tra le istituzioni pubbliche attraverso le reti telematiche, è auspicabile che, nel prossimo futuro, possano anche svilupparsi e diffondersi servizi di tipo informativo decisionale, in modo tale da rendere disponibile, all'interno di una rete sempre più ampia di pubbliche amministrazioni, il patrimonio informativo costruito nel tempo dalla Regione Piemonte. Da tutto ciò ne conseguiranno benefici per le amministrazioni stesse (ottimizzazione nell'utilizzo delle risorse, aumento di efficienza, integrazione nell'erogazione dei servizi, ecc.) e per il sistema sociale della regione (sportelli unici per servizi integrati, semplificazione amministrativa, distribuzione telematica dei servizi, ecc.).

Figura 5 - Architettura a n livelli logici per un Data Warehouse della Pubblica Amministrazione Piemontese



Maria Teresa Doriguzzi e Antonio Dal Borgo operano nella Divisione Registro Imprese di InfoCamere, società consortile delle Camere di Commercio, in qualità di progettisti di sistemi informatici destinati alla consultazione ed analisi delle informazioni relative alle anagrafiche delle imprese Italiane.

La relazione presenta l'esperienza del sistema delle Camere di Commercio nella progettazione di applicazioni in un'ottica di warehousing. Quindi evidenzia come sono state affrontate le tematiche di "Back-end" per il reperimento delle informazioni, la normalizzazione delle stesse, la successiva ristrutturazione nel Data Warehouse e in data base multidimensionali. Inoltre per la parte di 'front-end" vengono illustrate le caratteristiche dei prodotti :

- MOVIMPRESE (Natalità e mortalità delle imprese italiane registrate presso le Camere di Commercio)
- STATISTICHE ECONOMICO TERRITORIALI DELLE C.C.I.A.A.
- STATISTICHE DELL'OSSERVATORIO DEL COMMERCIO

Queste tre applicazioni insistono sul sistema informativo di maggior peso gestito dalle Camere di Commercio cioè il Registro delle Imprese Italiane, il Registro che, analogamente a quanto succede presso le anagrafiche Comunali per le persone fisiche, contiene la storia anagrafica di tutte le Imprese Italiane (nascita, morte, trasformazione, modifica ecc.) ed è costituito da 12.000.000 di soggetti.

Ogni giorno 6.000 terminali Camerali , un numero non facilmente quantificabile di utenti non Camerali (Banche, Assicurazioni ecc.) ed un numero sempre crescente di ricercatori consultano l'archivio per ottenere risposte a normali quesiti di Business.

Tra le richieste più frequenti troviamo la necessità di fornire sempre più velocemente e agevolmente informazioni statistiche sulla numerosità tipologia e andamento demografico delle imprese sul territorio nazionale.

Per far fronte a queste esigenze che di norma sono sempre diverse l'una dall'altra, si è ritenuto utile l'impiego di strumenti per semplificare il reperimento delle informazioni da un lato e la successiva rielaborazione locale dall'altro. Cioè non crediamo nel prodotto confezionato a priori bensì nel prodotto che il cliente si confeziona in funzione dell'analisi che intende effettuare. Ovviamente esistono dei semilavorati che possono essere trattati anche come prodotti finiti per semplificare al massimo l'operatività di chi è interessato ad ottenere informazioni immediatamente fruibili.

Ad esempio il progetto "Osservatorio del Commercio" commissionato ad InfoCamere dal Ministero dell'Industria edall'INDIS (organismo gestito daUnionCamere) prevede la fornitura di prospetti che successivamente e del tutto autonomamente vengono rielaborati dalle funzioni statistiche degli enti coinvolti. La produzione del

"Rapporto sugli aspetti strutturali del sistema distributivo italiano" redatto nel settembre 2000 dal Ministero dell'Industria ne è la prova tangibile.

Oggi l'informazione è richiesta in tempo reale, ovvero quello che è successo ieri non interessa più a nessuno, di conseguenza le applicazioni che intendono risolvere tali esigenze si scontrano con tre aspetti critici: le performance del sistema, la normalizzazione dei dati, la flessibilità di erogazione.

La grossa mole di dati e le elaborazioni a volte molto impegnative impongono architetture complesse che vanno dalla acquisizione di sistemi specifici per la gestione delle memorie di massa, alla elaborazione parallela, allascalabilità dei sistemi per l'erogazione del servizio on-line.

Fondamentale per l'analisi di archivi amministrativi è una facile manutenzione della normalizzazione dei dati da cui poter ricavare informazioni economico statistiche.

Un esempio è la pubblicazione trimestrale di "Movimprese", realizzata dal sistema delle Camere di Commercio, in cui è evidente la trasformazione dell'archivio amministrativo del Registro delle Imprese in una rilevazione statistica.
Il prodotto rileva la numerosità e la distribuzione sul territorio (provinciale, regionale e nazionale) di tutti i soggetti economici tenuti all'iscrizione presso il Registro delle Imprese delle Camere di commercio Italiane (a regime dal 27/01/1997); di seguito una breve descrizione delle caratteristiche:

I PARTE (economico-statistica): panoramica del sistema imprenditoriale italiano

⇨ andamento demografico delle imprese (imprese, unità locali e cariche)

⇨ continuità con il Registro Ditte

II PARTE (statistico-amministrativa): strutturagiuridico-amministrativa delR.I.

⇨ le dimensioni amministrative del Registro delle Imprese

⇨ SEZIONI (sezione ordinaria e 4 sezioni speciali ed il REA)

– Analizza i fenomeni demografici per settore di attività economica e per tipologia di forma giuridica dell'impresa

Variabili stock:

Imprese Registrate, Imprese Attive

Variabili flusso:

ISCRIZIONI, CESSAZIONI, VARIAZIONI

Territorio

Provincie, Regioni

Settori di attività economica

Sezioni (1 lettera), sottosezioni (2 lettere) e divisioni (2 cifre) di attività

Tipologie di forma giuridica:

Società di capitale, Società di persone, Ditte individuali, Altre forme

- Fornisce informazioni socio-economiche di enorme rilievo: consente di censire trimestralmente categorie come quelle dei **piccoli imprenditori**, **imprenditori agricoli**, **coltivatori diretti** ed **artigiani**.
- Costituisce un importante capitolo dell'informazione statistica al servizio degli operatori economici e dei "decision maker" italiani
- I dati sono disponibili all'indirizzo Internetwww.infocamere.it

Per incrementare la fruibilità e la flessibilità di questi servizi, è stato realizzato uno strumento, inizialmente messo solo a disposizione della rete delle Camere di Commercio, basato sull'impiego di data base multidimensionali che consentono l'esplorazione delle informazioni in una modalità decisamente più dinamica.
In sintesi è l'utilizzatore del sistema che decide le variabili di analisi , quelle di classificazione, la gerarchia di navigazione, la profondità di esplorazione ed i filtri da applicare per concentrare l'analisi.

Il sistema Camerale oggi, utilizzando questi sistemi, incrocia agevolmente ad esempio le informazioni territoriali dell'impresa con quelle sullo stato di nascita dell'imprenditore, per approfondire i fenomeni di imprenditoria straniera in Italia.

In realtà i nostri progetti impongono qualche limite all'utenza in quanto non è possibile rendere navigabile , in tempi accettabili, qualsiasi tipo di incrocio.
All'aumentare delle informazioni e della granularità delle stesse aumenta a dismisura la necessità di spazio disco e di conseguenza la capacità e rapidità della CPU.

Oltre certi limiti il rapporto costo beneficio rende di fatto preferibile il ricorso a strumenti più tradizionali quali la programmazione per risolvere i problemi specifici ricadendo potenzialmente nei classici problemi di manutenzione ed adeguamento del software.

Nel progettare questi sistemi InfoCamere non aveva del tutto sotto controllo gli aspetti precedentemente discussi e la strumentazione relativamente nuova non ha consentito di verificare esperienze analoghe, ma la filosofia di queste tecnologie basata sulla gestione di metadati piuttosto che nella scrittura di software proprietario ha stimolato la sperimentazione di opportune soluzioni.

L'esito positivo di quest'ultime, sia sotto l'aspetto della qualità del prodotto realizzato sia sotto quello della compressione significativa dei costi e dei tempi, ha aperto la strada alle realizzazioni precedentemente citate che hanno lo scopo di diffondere le potenzialità e la conoscenza delle informazioni contenute negli archivi gestiti dalle Camere di Commercio.

Quinta
Conferenza
Nazionale
di Statistica

WORKSHOP

Sistemi informativi territoriali

coordinatore

Cristoforo S. Bertuglia
docente del Politecnico di Torino

Roma 15, 16, 17 novembre 2000
Palazzo dei Congressi
Piazzale Kennedy - EUR

ISTAT

WORKSHOP

Sistemi informativi territoriali

coordinatore

Cristoforo S. Bertuglia
docente del Politecnico di Torino

SOMMARIO

G. Biallo	Il mercato delle informazioni geostatistiche: le esigenze degli enti pubblici e le necessità dei privati
S. Occielli	Verso un nuovo modello di sistema informativo territoriale (Un sistema informativo in azione)
A. Spaziante e P. Cipriano	Valorizzare il patrimonio informativo di SIT locali
F. Vico	Le infrastrutture dell'informazione geografica: cooperazione e condivisione dei dati geografici di base

WORKSHOP

Sistemi informativi territoriali

Il mercato delle informazioni geostatistiche: le esigenze degli enti pubblici e le necessità dei privati

G. Biallo

Il mercato delle informazioni geostatistiche: le esigenze degli enti pubblici e le necessitá dei privati

Giovanni Biallo

Università degli Studi "La Sapienza" di Roma - Facoltà di Architettura - Diploma Universitario sui Sistemi Informativi Territoriali

Sommario

L'analisi dei rapporti tra produttori e fruitori di datigeostatistici porta inevitabilmente ad evidenziare le idiosincrasie tra le tecnologie proposte dai produttori e i giacimenti informativi disponibili a livello territoriale. A partire dall'approccio seguito dall'Istat per la diffusione delle basi territoriali del Censimento del 1991 si è innescata una scintilla che ha portato ad oltrepassare i confini delle tematizzazioni classiche ed a sfruttare le tecniche combinate di analisi spaziale e statistica, il cui punto di incontro è il fulcro delle potenzialità future dei SIT.

1. Il mercato del GIS

Nato con la carta geografica, primo esempio di sistema informativo territoriale, il GIS (Geographical Information System) oggi identifica le tecnologie informatiche per l'elaborazione di dati geografici. Con l'avvento degli elaboratori elettronici e le nuove idee sull'analisi spaziale, agli inizi degli anni '70 nel Nord America videro la luce i primi software GIS e le prime applicazioni in settori quali la pianificazione territoriale e l'analisi di datigeostatistici.
Sbarcato molto presto in Europa (Gran Bretagna, Svezia, Finlandia, Germania e Francia), verso la metà degli anni '80 il GIS giunge finalmente in Italia. I primi a fornirsi di sistemi e in contemporanea a produrre dati territoriali in formato numerico sono stati gli enti pubblici: Ministeri, Servizi Tecnici dello Stato, Amministrazioni Regionali. La corsa all'informatizzazione geografica ha investito in breve tempo molti altri enti pubblici ed anche società private. Oggi, dopo un periodo di stasi legato alla caduta della prima repubblica ed alle difficoltà che hanno investito l'economia italiana, il GIS affronta il mercato in maniera molto più determinata verso un'utenza più smaliziata ed attenta alle scelte tecnologiche.
Il mercato italiano vede da un lato la resistenza degli enti centrali a produrre cartografia numerica e a renderla di pubblico dominio, dall'altro la disponibilità delle amministrazioni regionali e provinciali a produrre e diffondere banche dati geografiche locali. Questa dicotomia genera una disomogeneità nella disponibilità di informazioni e nei prezzi di acquisto che spesso determina incertezza sull'investimento che un utente deve sostenere per dotarsi di un Sistema Informativo Territoriale.
Sulla disponibilità di software GIS oggi osserviamo una spinta verso una forte diversificazione di prodotti progettati per esigenze specifiche: prodotti per la consultazione dei dati, per la gestione di grandi quantità di dati, per lo sviluppo di interfacce specifiche, per la consultazione via Internet, ecc. Sicuramente la tecnologia hardware più seguita dai produttori è basata su elaboratori con processori Intel su sistemi Microsoft Windows NT.
Il mercato è animato anche da una forte presenza di società che si sono specializzate nello sviluppo di applicativi verticali basati su prodotti affermati. Ma sono anche molte le aziende che producono software GIS "made in Italy" molto efficienti e capaci di interagire con i software più blasonati. Le

applicazioni ormai spaziano in moltissimi settori più o meno tradizionali per il GIS: urbanistica, pianificazione territoriale, ambiente, beni culturali, traffico, viabilità, trasporti, reti tecnologiche, facility management, geomarketing, statistica, telecomunicazioni, ecc.
Il quadro del mercato italiano si completa con gli utenti che sono al centro delle nostre attenzioni. Nella Pubblica Amministrazione Centrale la presenza di applicazioni GIS è abbastanza disuniforme e dipende essenzialmente dalla spinta più o meno innovativa che ogni ministro vuole dare al proprio ministero. Ovviamente con i continui cambi al vertice, con le inefficienze burocratiche e di gestione dei fondi, i ministeri non sono certo un esempio da perseguire. Per contro le regioni e le province del centro-nord sono nella maggior parte dei casi già fornite di sistemi e dati GIS; stentano invece a partire quelle del sud tranne rarissime eccezioni. I comuni sono invece la nuova utenza su cui puntare. Il GIS, fino a qualche anno fa utilizzato solo da alcuni grandi comuni, oggi, grazie ai prezzi più abbordabili alla disponibilità di applicativi verticali e di banche dati geografiche, vede una forte espandibilità all'interno delle amministrazioni comunali.
Il nocciolo della questione però, non è rappresentato dagli strumenti o dai dati ma, dalla grande difficoltà di inserimento operativo del GIS nelle procedure tecnico-amministrative e di riqualificazione del personale in organico al fine di un uso effettivo della tecnologia.
Spesso un progetto di integrazione di soluzioni GIS nelle procedure tradizionali non viene eseguito correttamente per mancanza di una attenta analisi costi-benefici che potrebbe garantire l'individuazione di obiettivi concreti e di relativi benefici tangibili.
Più rapido l'inserimento della tecnologia GIS nelle società private che fanno molta più attenzione della pubblica amministrazione ai benefici. Settori in forte espansione vedono le applicazioni per il geomarketing o per le telecomunicazioni fra le più gettonate sul mercato.
Il mercato italiano è oggi considerato dai fornitori di software GIS molto interessante ed è considerato come un ottimo terreno di prova per definire prodotti e piani di sviluppo applicabili in seguito su tutto il mercato europeo. Per questo motivo molti fornitori si stanno affacciando direttamente e senza interposti soggetti sul nostro mercato.
Con la presenza diretta dei fornitori software, il supporto di società di servizi specializzate nello sviluppo di applicativi verticali, una utenza che grazie anche al supporto dell'editoria specializzata è molto più preparata ed attenta alle evoluzioni tecnologiche, ed uno sforzo verso una maggiore liberalizzazione dei dati geografici, si potrebbe configurare un mercato ancora più interessante ed esteso che identifica il GIS non più come una componente di un sistema informativo generico ma come un sistema di integrazione e diffusione di informazioni sul territorio.

2. Il mercato dei dati geostatistici e l'esempio americano

Per avere un previsione dell'evoluzione del mercato dell'informazione geografica in Italia, è senz'altro utile osservare l'evoluzione della produzione dei dati negli Stati Uniti, patria dei sistemi GIS. Sin dagli anni 70 gli enti governativi americani si sono impegnati in modo determinato nella produzione di dati geografici che sono stati subito messi a disposizione del pubblico con il concetto che l'informazione prodotta dal governo è di pubblico dominio ed il cittadino deve poterne usufruire senza ulteriori costi, poiché ha già contribuito alle spese di produzione pagando le tasse. In prima linea troviamo la Defence Map Agency, il Geologycal Survey, il Census Bureau. Quest'ultimo si è subito preoccupato di produrre le coperture geografiche che hanno un diretto riferimento con i dati statistici: limiti amministrativi a vari livelli e sezioni di censimento. Oltre a queste basi ha prodotto un grafo stradale completo di toponomastica di tutto il territorio degli Stati Uniti. Le due banche dati sono risultate le più utilizzate in assoluto dagli utenti americani perché con esse è possibile georeferenziare non solo l'informazione statistica ma anche tutte quelle banche dati alfanumeriche che hanno come georiferimento un indirizzo postale. Una politica rivolta non solo a diffondere a basso costo il dato ma anche ad autorizzare le aziende private alla rivendita del dato integrato da

nuove informazioni di varia natura, ha aiutato ad ottenere la massima diffusione dell'informazione e soprattutto l'utilizzo degli strumenti software GIS.

La disponibilità di questi dati sin dalla fine degli anni settanta, oltre a dare un valido supporto agli enti pubblici centrali e locali e ai pianificatori per la comprensione delle evoluzioni antropiche del territorio, ha incrementato di fatto la nascita di aziende specializzate nel geomarketing, cioè quella nuova metodologia che utilizza non solo l'informazione geografica ma anche e soprattutto le funzioni di analisi spaziale per ottimizzare i modelli di analisi di marketing.

In una mia visita ad alcune di queste aziende nei primi anni novanta ho potuto constatare che le richieste del vasto mercato americano erano tali da richiedere centinaia di specialisti e grossi investimenti per la costituzione di banche dati sempre più dettagliate ed aggiornate.

3. Il mercato italiano

Come negli Stati Uniti, se pur con il tradizionale ritardo cronico, anche in Italia l'evoluzione dei dati geostatistici e delle relative applicazioni ha seguito lo stesso percorso. Alla fine degli anni 80l'Enel ha stipulato una convenzione con l'Istat per la produzione in formato numerico dei limiti amministrativi comunali tratti dalle carte IGMI e aggiornati in occasione del Censimento del 1981. Con il Censimento del 1991 l'Istat è partito con un grande progetto con il quale realizza le basi territoriali composte dai limiti amministrativi, località abitate e sezioni di censimento. La base dati è a tutt'oggi la più diffusa in assoluto. Oltre ai comuni, che hanno ricevuto copia dall'Istituto, è stata fornita e venduta ad enti pubblici e società private ed è tuttora la banca dati più richiesta in assoluto nonostante i dati statistici associati sono ormai storici. Mentre però i dati geostatistici siffatti sono utilizzati in modo determinante soprattutto per la pianificazione sia urbanistica che di area vasta, non lo sono stati nel settore privato per le analisi di geomarketing. Uno dei limiti che ha determinato questa scarsa diffusione applicativa l'ho potuto riscontrare in un seminario di geomarketing organizzato a Roma nel 1997 dalla rivista MondoGIS: quasi tutte le applicazioni ed i prodotti presentati erano finalizzati alla rappresentazione ed alla consultazione in forma geografica del dato. In pratica si riscontra la presenza da un lato di aziende che provengono dai settori tradizionali del GIS è tentano di fare del geomarkenting, utilizzando anche le funzioni di analisi spaziale (overlay, buffering, network, ecc.) ma non avendo alcuna conoscenza di marketing, e dall'altro di aziende specializzate in analisi di marketing che usano gli strumenti GIS solo per le funzioni di rappresentazione geografica del dato statistico. Oggi in Italia le aziende che riescono a realizzare dei veri servizi di geomarketing si possono contare sulle dita di una mano e sono anche le uniche che possiedono banche dati geostatistiche molto interessanti anche se non accessibili al pubblico.

Il confronto con il settore degli enti che si occupano di analisi statistica è presto fatto: quei pochi che usano gli strumenti GIS li applicano quasi sempre solo per la rappresentazione geografica di dati elaborati con strumenti statistici (SAS, SPSS, ecc.) non usufruendo delle notevoli potenzialità di analisi geografica che gli strumenti GIS hanno. La combinazione di software per l'analisi statistica e di software per l'analisi geografica porterebbe a risultati decisamente più interessanti ed affidabili.

La formazione sull'uso integrato degli strumenti software a partire dall'Università per finire ai pianificatori, statistici, specialisti di marketing ed esperti di GIS, è alla base della soluzione del problema.

4. L'integrazione delle banche dati geografiche e alfanumeriche

Il Census Bureau americano sin dall'inizio aveva inquadrato il problema: le basi geografiche per georeferenziare informazioni statistiche sono due: le sezioni di censimento ed il grafo stradale.

L'integrazione delle due, porta un valore aggiunto ineguagliabile permettendo di generare nuovi dati statistici e di aggiornare quelli già esistenti. In pratica, pensate quante banche dati alfanumeriche contengono come elemento di georiferimento sul territorio un indirizzo postale (via, numero civico, CAP, località). Per esempio i dati contenuti nelle anagrafi comunali permettono digeoriferire la popolazione sull'arco di strada sul quale insiste l'abitazione abituale. Sommando per classi questa popolazione si possono eseguire analisi spaziali utilizzando il grafo, per esempio nel settore dei trasporti pubblici. Se esiste un legame tra sezione di censimento ed indirizzi postali si possono travasare i dati delle anagrafi, per quantità e classi, nelle sezioni, aggiornando in tal modo alcuni indicatori sulla popolazione che i Censimenti ci danno ogni dieci anni. La migrazione di dati fra sezioni di censimento e archi di strada permette inoltre di usufruire di un numero di modelli di analisi spaziale molto ampio, garantendo all'utente l'utilizzo dei medesimi dati statistici.

L'Istat ha progettato le basi territoriali dei Censimenti 2000/2001 con un occhio di riguardo a questa problematica e sta pensando, insieme ad altri enti direttamente interessati all'argomento, a definire degli standard per la normalizzazione dei parametri di georeferenziazione dell'informazione.

Gli enti locali ed i pianificatori che già si sono scontrati con questa problematica attendono con ansia la disponibilità di una base dati integrata che gli garantirà digeoriferire le banche dati alfanumeriche in proprio possesso, di generare nuova informazione statistica e di verificare l'evoluzione antropica del territorio.

Bibliografia

Aronoff (1989), "Geographic Information Systems: a management perspective", Ottawa, Canada, WDL Pubblication.

Autori Vari (2000), "Oltre i Confini del GIS", Atti della Seconda Conferenza di MondoGIS, Roma, MondoGIS.

Autori Vari (1997), "La progettazione dei Censimenti 1991: basi territoriali, organizzazione della rete di rilevazione, campagna di informazione, piano dei controlli", Roma, Istat.

Biallo (1996), "Dalla geografia al GIS: venticinque secoli di storia", in MondoGIS n. 1, Milano, Ziviani Editore.

Biallo, Consorti (1996), "La Standardizzazione Europea nell'Informazione Geografica: Situazione Attuale e Sviluppi Futuri", in Atti del Settima Conferenza Nazionale AM/FM GIS, Pisa.

Maguire, Goodchild, Rhind (1991), "Geographical Information Systems", England, Longman Scientific & Technical.

Vico (1996), "GIS e Pianificazione Urbanistica", Segrate (MI), Editrice Il Rostro.

WORKSHOP

Sistemi informativi territoriali

Verso un nuovo modello di sistema informativo territoriale (Un sistema informativo in azione)

S. Occielli

Verso un nuovo modello di sistema informativo territoriale
(Un sistema informativo in azione)

Sylvie Occelli

IRES - Istituto di Ricerche Economico Sociali del Piemonte, Via Nizza 18, 10125 Torino
e-mail:occelli@ires.Piemonte.it

SOMMARIO

L'affermarsi di una società sempre più imperniata sull'informazione, la diffusione delle nuove tecnologie di comunicazione, il riconoscimento della crescente responsabilità dell'azione individuale rappresentano alcuni dei processi di cambiamento che stanno investendo i sistemi urbani alle diverse scale territoriali.

Essi hanno delle profonde implicazioni sulla natura delle 'informazioni' territoriali oggi richieste e/o auspicate; sia con riferimento alla domanda di informazione (tipo di informazione, scala territoriale, frequenza temporale), sia in termini di produzione/gestione di informazione (ambiente informativo, intra-net ed internet, pacchetti GIS, approccio metodologico utilizzato), sia dal punto di vista della comunicazione delle informazioni stesse (trasmissione/scambi di informazione tra soggetti diversi), nonché, infine, dal punto di vista dell'accessibilità da parte dei diversi utenti o dei (sempre più numerosi) fruitori (barriere di accesso, restrizione di uso, vincoli di privacy).

Una conseguenza generale è che, sia dal lato sia delle aspettative, sia da quello delle necessità, ci si trova di fronte a situazioni mutevoli nel tempo, nelle quali, il più delle volte, si riscontra una carenza pregressa di informazioni, ma in altre si può essere in presenza di un sovraccarico informativo, che genera confusione. Più spesso ancora ci si trova di fronte a situazioni in cui la domanda di informazioni si caratterizza per livelli di complessità, che stenta a trovare una risposta adeguata nei dati disponibili.

Tutto ciò influisce in misura significativa sul 'modello' di 'sistema informativo territoriale' prefigurabile o che sarebbe opportuno realizzare nella realtà.
Anche se non facilmente delineabile, un fatto chiaramente percepito è che esso si discosta, in misura considerevole, da quello che stava alla base delle architetture dei sistemi informativi territoriali sviluppati fino a pochi anni fa.

Riflettere sui requisiti attesi o auspicabili di un tale modello appare, pertanto, un esercizio meritevole di interesse. In questa direzione, la presente nota si propone di illustrarne alcuni, a partire da una riflessione critica in ordine ad alcune caratteristiche del modello tradizionale.

1. Introduzione

I sistemi informativi territoriali rappresentano un filone di lavoro che ha suscitato, soprattutto a partire degli anni '80, un interesse crescente da parte di un pubblico sempre più vasto ed eterogeneo formato non solo da studiosi del mondo accademico ma anche da professionisti, da funzionari della pubblica amministrazione e da cittadini comuni (Nijkamp e Rietveld, eds., 1984, Baherenberg, Fischer e Nijkamp eds, 1984, Van Geenhuizen e Nijlamp, 1997)

Essi nascono da un felice connubio tra un'esigenza diffusa di disporre di approcci efficaci per il trattamento delle informazioni ai fini della pianificazione e della gestione dei fenomeni territoriali e le crescenti potenzialità di calcolo rese disponibili dallo sviluppo dei calcolatori e delle nuove tecnologie delle comunicazioni. Il progresso tecnico, in particolare, rappresenta il fattore decisivo che maggiormente ha influito sulla fortuna e sulla diffusione dei sistemi informativi territoriali (SIT). Qualunque sia l'architettura che sta alla base di un SIT, essa presuppone, di fatto, un qualche sistema di calcolo, in grado di archiviare, elaborare ed interrogare informazioni, corredato, eventualmente, anche dalla possibilità di restituirle visivamente tramite opportuni tematismi grafici.

In realtà i SIT sono entità ibride definite dall'interazione di tre principali componenti:

- Una visione del territorio
- Una strumentazione tecnologica
- Un insieme di utilizzatori.

Proprio tale natura ibrida costituisce una peculiarità dei SIT e, forse, la principale ragione del loro successo. In passato, le tre citate componenti erano ricondotte ad una concezione unitaria sulla quale si imperniava l'architettura generale dei SIT. La componente 'strumento' (grazie alle potenzialità tecnologiche) era concepita e realizzata in modo tale da soddisfare le domande di informazione poste dagli utilizzatori in ordine a certi problemi di analisi e/o di controllo dei fenomeni territoriali. Questi problemi, a loro volta, erano opportunamente, scomposti, articolati e categorizzati in modo da trovare risposte soddisfacenti nella componente 'strumento'.

Oggi tale unitarietà sta venendo meno e le componenti citate appaiono sempre più autonome. In questo senso si può affermare che la concezione dei SIT stia subendo un'evoluzione.

Le ragioni sono molteplici e trovano una spiegazione, da un lato, nei processi di cambiamento nella città e, dall'altro, nei modi stessi di studiare i fenomeni ed i processi urbani.

Tutto ciò influisce in misura significativa sul 'modello' di 'sistema informativo territoriale' prefigurabile o che sarebbe utile predisporre. Per quanto non facilmente delineabile, un fatto chiaramente percepito è che il nuovo modello si discosta, in misura considerevole, da quello che stava alla base delle architetture dei sistemi informativi territoriali sviluppati fino a pochi anni or sono.

Queste note si propongono di sviluppare alcune considerazioni sui requisiti attesi o auspicabili del nuovo modello. In questa direzione, il paragrafo 2 richiama i principali aspetti di cambiamento in atto che stanno avendo un'influenza significativa sulla concezione dei SIT e nel paragrafo 3 se ne evidenziano alcune conseguenze. Alcune considerazioni generali sulle prospettive di sviluppo futuro dei SIT concludono il lavoro.

2. Fattori di cambiamento nella concezione dei SIT

In una società in rapida evoluzione quale quella attuale, vi sono svariati e molteplici fattori di cambiamento che sollecitano un ripensamento nella concezione generale dei SIT. Per quanto difficilmente descrivibili in modo esaustivo in poche righe, essi possono tuttavia essere ricondotte a tre principali tipi (Occelli, 1999) :

a) Modificazioni delle aspettative conoscitive;
b) Diffusione dell'innovazione tecnologica;
c) Cambiamenti nel contesto socio-culturale (mutamenti negli utilizzatori dei SIT e trasformazioni del 'tipo di questioni' oggetto di investigazione).

a) Modificazioni delle aspettative conoscitive

Un primo ed importante tipo di cambiamento riguarda le aspettative conoscitive o, in altre parole, la natura stessa delle conoscenza richieste (Boulding, 1985, Maestre, 1994). Tali cambiamenti possono essere colti da diversi punti di vista, ed in particolare :

- Dal punto di vista dell'evoluzione del pensiero scientifico. Gli avanzamenti significativi riguardano, da un lato, il superamento di una visione ingenua del rapporto tra osservatore e realtà di osservazione e, dall'altro il riconoscimento che la realtà dei processi sociali è, per sua natura, dinamica, mutevole ed incerta. Una conseguenza di rilievo concerne in particolare il passaggio da un atteggiamento che guarda alla realtà come qualcosa di fisso e di immutabile, ad uno che guarda alla realtà come entità in continua evoluzione, che richiede pertanto una revisione /adeguamento continuo dei modi di osservare e di intervenire sulla realtà stessa (Turoff, 1997).

- Dal punto di vista della matrice disciplinare che legittima una 'conoscenza scientifica'. A questo proposito, l'aspetto forse più interessante del dibattito in corso è costituito dalla 'tensione' tra il consolidamento del proprio campo disciplinare (che risponde ad una necessità più generale di 'normalizzazione' dei propri paradigmi) ed un'esigenza diffusa di adeguamento al 'paradigma della complessità' (Bertuglia, Bianchi e Mela eds., 1998, Casti, 1997) e della sua finalizzazione all'interno di ciascun campo di studio.

- Dal punto di vista, infine, dei rapporti tra l'esperto (la comunità scientifica) e gli altri gruppi sociali. Si va affermando una visione più laica della comunità scientifica, che sempre più spesso è chiamata ad interagire con le altre comunità che fanno parte della collettività sociale. Ne consegue che rilevanza crescente assumono, da un lato, la definizione dei 'contenuti di utilità sociale' che possono essere invocati per giustificare un'attività conoscitiva e, dall'altro, la comunicazione degli esiti di tale attività fra le diverse comunità.

b) L'innovazione tecnologica

Lo sviluppo delle tecnologie informative e la diffusione dei nuovi mezzi di comunicazione, hanno avuto, soprattutto nell'ultimo decennio, un impatto enorme sulla dotazione della 'strumentazione metodologica' utilizzabile per l'analisi dei processi territoriali (Bertuglia, Occelli e Lombardo, 1998, Cecchini, 1999, Maciocco a cura di, 1994, Occelli e Rabino, 2000). Essi configurano, anche per queste analisi, un vero e proprio 'technological backcloth', le cui possibilità vanno al di là della costruzione di architetture informative sempre più sofisticate e potenti (quali prefigurabili nella realizzazione di sistemi informativi' che interfacciano tecniche di rappresentazioni GIS, con moduli di simulazione, moduli per la valutazione, moduli per il data-mining ecc.).

Tali possibilità sono molteplici ed ancora largamente inesplorate. Alcune di loro riguardano, ad esempio:

- Il cambiamento del ruolo del computer. Se, in passato, il computer era semplicemente lo strumento che rendeva possibili certe operazioni computazionali (il cui 'successo' era condizionato in ultima istanza dal grado di sofisticazione tecnologica), oggi il computer diventa parte integrante dell'applicazione metodologica (senza il computer cioè, il metodo non sarebbe applicabile).

- L'aumento di efficienza nell'uso del metodo. Oltre alla possibilità di trattare dimensioni di problemi precedentemente inimmaginabili, lo sviluppo di nuovi 'media' rende possibile introdurre significativi elementi di innovazione nell'applicazione di approcci di analisi tradizionali (si pensi, ad esempio, all'applicazione della computer vision all'analisi multivariata).

- Il potenziamento delle possibilità di comunicazione. L'adozione delle nuove tecnologie informative presso un pubblico sempre più ampio ed eterogeneo, contribuisce a far superare quell'atteggiamento 'anti-scientifico' lamentato da alcuni studiosi alcuni anni or sono. La possibilità inoltre di 'sperimentare', in modo autonomo, allarga le prospettive applicative, permettendo di condividere le esperienze fatte da altri e di realizzarne di proprie. Da questo punto di vista, l'uso di metodi, si rivela un supporto fondamentale di apprendimento circa il 'come conoscere' i fenomeni socioeconomici che ci circondano.

c) I cambiamenti del contesto socio-culturale

Un'ultima ma non meno importante fonte di cambiamento riguarda il 'contesto socio-culturale'. Oltre ad un innalzamento del livello di informazione della società, si assiste ad un aumento della domanda di conoscenza circa i fenomeni socioeconomici, la quale tende a diventare sempre più articolata e selettiva (Andersson et Al. eds., 1993, Castells, 1989).

Esso può ritenersi l'esito di una più generale tendenza 'di emancipazione' della società occidentale, alla quale contribuiscono svariati processi quali, in particolare, quelli relativi a: a) lo sviluppo della personalità umana, b) l'affermazione di una tendenza 'epocale' verso il decentramento che, a sua volta, presenta aspetti di natura economica, organizzativa ed istituzionale (Bennet, 1996), c) la crescente sensibilità alle questioni della sostenibilità dello sviluppo e d) la (già menzionata) pervasività dell'impatto delle nuove tecnologie informative (Bertuglia e Occelli, 1995, Couclelis, ed., 1996, Graham e Marvin, 1996).

Nel dibattito in corso circa la transizione ad una società post-industriale, cresce pertanto la consapevolezza circa il fatto che l'evoluzione ed i cambiamenti dei sistemi socioeconomici e territoriali dipendono da una molteplicità di comportamenti, individuali e collettivi, i cui effetti si dispiegano su diverse scale spazio-temporali, determinando una gamma di possibili sentieri di evoluzione.

Cresce inoltre la consapevolezza che affrontare il cambiamento non sia solo una necessità non eludibile, dettata da esigenze di aggiustamento continuo verso posizioni di stabilità e di certezza. Infatti, si va affermando la convinzione che, per sopravvivere in un mondo in trasformazione, oggi occorra una maggiore attenzione a cosa si sta facendo: alle azioni, alle loro modalità' di realizzazione ed ai loro effetti sull'ambiente circostante.

In questo quadro, anche le 'situazioni di criticità' si presentano sotto una nuova luce. Oltre all'esigenza di definire i 'problemi' che devono essere oggetto di intervento da parte della politiche, emergono nuove esigenze di formulare 'politichepro-attive', capaci non solo di porre rimedio, ma, anche, di 'anticipare' eventuali situazioni critiche.

3. Nuovi aspetti nella concezione dei SIT

Come introdotto in precedenza, un SIT è costituito da tre principali componenti, Fig.1:

- Una visione del territorio
- Una strumentazione tecnologica
- Un insieme di utilizzatori.

Insieme, queste componenti concorrono a determinare il principale output di un SIT: l'informazione. Se prescindiamo dalla sua definizione in termini matematici, il concetto di informazione non è definibile in modo univoco, ma può essere declinato lungo un continuo che va dal semplice fatto ad una qualche forma di più elaborata di conoscenza. Una definizione generale, particolarmente adatta alla discussione presente, è quella secondo la quale, l'informazione può essere considerata 'come l'esito di una manipolazione creativa di dati e di saperi precedenti, finalizzata ad aiutarci a raggiungere nuovi livelli di comprensione ' (Janelle e Hodge, 2000).

La concezione di SIT che è evidenziata in tale figura, pertanto, non è riconducibile semplicemente alla predisposizione di un prodotto informatico, ma è assai più ampia. Ciò è stato già da tempo riconosciuto da diversi studiosi (Maciocco a cura di, 1994). Las Casas e Storchi, (1994, p. 422.), ad esempio, argomentano che: 'La accezione allargata di Sistema Informativo include quelle informazioni attorno alle quali il processo di decisione si sviluppa: fabbisogni e aspirazioni, obiettivi e priorità relative. Alternative e outcomes di tali alternative'

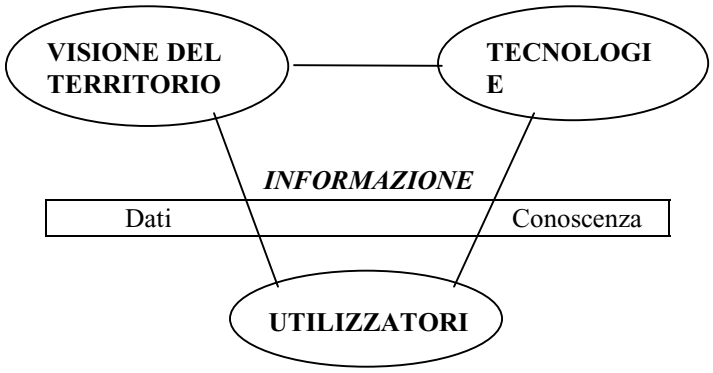


Figura 1. **Le principali componenti di un SIT**

Per quanto sia proprio tale accezione più ampia dei SIT che maggiormente risente dei cambiamenti richiamati nel paragrafo precedente, vale la pena soffermarsi brevemente su alcune delle conseguenze che si osservano per le singole componenti dei SIT

A questo scopo è stata predisposta la Tab.1, la quale (pur con ovvie semplificazione) evidenzia le principali caratteristiche delle componenti in una concezione tradizionale dei SIT (modello tradizionale) ed in una che, per effetto di quei cambiamenti, se ne discosta (nuovo modello). Va osservato, per inciso, che le caratteristiche del nuovo modello non annullano quelle che esistevano del modello tradizionale, ma ne costituiscono un arricchimento od un'evoluzione. E' questo il caso, in particolare, di alcune specifiche dei SIT relative al 'ruolo' ed alle 'funzioni', le cui caratteristiche rilevate nel nuovo modello non potrebbero esistere senza quelle previste nel modello tradizionale.

Tabella 1 **Caratteri salienti della concezione dei SIT nel modello tradizionale ed in quello nuovo**

	Modello tradizionale	**Nuovo modello**
Visione del territorio		
natura dei fenomeni	multidimensionalità e molteplicità	diversificazione, evoluzione, articolazione multi-livello
modi di investigazione	approcci volti all'interpretazione ed alla spiegazione	approcci volti alla comprensione
Strumentazione tecnologica	nodo autonomo di elaborazione	rete di nodi caratterizzati da diversa specializzazione e capacità computazionali
Utilizzatori	pochi ed esperti	numerosi, con diversa expertize anche non esperti
Specifiche del SIT		
Ruolo	sistema di supporto alle decisioni	ambiente culturale delle 'meta-conoscenze' e dell'esperienza atto a favorire l'azione
Funzione	monitoraggio, valutazione	strutturazione e anticipazione di problemi
Principali finalità del trattamento dei dati	registrazione, validazione, classificazione, archiviazione, manipolazione (riduzione delle dimensionalità, estrapolazione, ecc.)	esplorazione, simulazione, comunicazione
Informazione attesa	conoscenza oggettiva e procedurale	conoscenza incorporata

A) Con riferimento alla visione del territorio e specificatamente al campo della geografia quantitativa, l'impulso forse maggiormente significativo che è derivato dai cambiamenti sopra richiamati riguarda l'introduzione di nuovi modi diconcettualizzare i fenomeni ed i processi che governano le dinamiche dei sistemi urbani (Bertuglia, Bianchi e Mela eds., 1998, Fotheringham e Brunsdon, 1999). Essi riposano sul riconoscimento che:

- le interdipendenze presenti nel sistema urbano sono all'origine di non-linearità, suscettibili di produrre discontinuità nella traiettoria evolutiva del sistema stesso (criticità, fenomeni catastrofici, biforcazioni) ed effetti inattesi o contro-intuitivi;
- i cambiamenti urbani sono l'esito di aggiustamenti continui, che derivano dai comportamenti dei diversi sottosistemi (attori urbani, ambiente esterno) in interazione tra loro. Fenomeni di auto-organizzazione possono prodursi e dare luogo a modificazioni qualitative che cambiano, anche radicalmente, la struttura dei sistemi urbani (Lombardo, 1994);

- l'equilibrio non è uno stato definitivo ma una situazione transitoria, contrassegnata, per sua natura, da instabilità e da oscillazioni;
- cambiamenti, aggiustamenti e situazioni di equilibrio si specificano, attraverso un'articolazione locale-globale, nella quale l'articolazione territoriale presenta rilevanza crescente.

Un tale approccio di analisi nei confronti dei sistemi territoriali si accompagna inoltre alla consapevolezza che esiste un divario tra il mondo 'descritto' - e cioè le nostre rappresentazioni del mondo - ed il mondo reale o percepito che si desidera analizzare che è causa di errore ed incertezza[1].
L'aspetto forse più rilevante da sottolineare, nel nuovo modello di SIT, riguarda pertanto il riconoscimento esplicito che la 'complessità' dei fenomeni territoriali non può essere affrontata privilegiando un solo ed unico approccio: un approccio, cioè, che si preoccupi di distillare una conoscenza precisa, univoca, essenziale, oggettiva da una qualche grande mole di dati.

Ciò significa, come argomentato ad esempio da Casetti (1999), che diversi modi di investigazione (quelli che l'autore chiama modo letterario, modo accademico e modo matematico) sono necessari e devono essere opportunamente tenuti in considerazione e messi in relazione fra loro[2].
Per meglio illustrare questo punto può essere utile considerare il tipo problema associato ad un certo fenomeno territoriale, classificandolo in termini di cause e di effetti prodotti, come ad esempio proposto da Minsky (2000), Tab.2.

'….Quando le cause sono poche e gli effetti modesti, (celle a in alto a sinistra) il problema sarà facile da risolvere (tramite opportune procedure di interrogazione o certi algoritmi di ricerca). In presenza

---

[1] Il termine errore, ad esempio, può essere interpretato in diversi modi. Comunemente, esso sta ad indicare uno scollamento tra la realtà e la nostra rappresentazione della realtà (Unwin, 1995), in particolare:

1. sottende una nozione di 'accuratezza, ovvero di 'prossimità' di certi valori, calcolati o stimati, rispetto a certi valori di riferimento, considerati come giusti. Oltre ad una nozione di precisione (quale espresso dal numero di cifre significative in una misurazione), ciò richiama anche un concetto di giudizio di valore. Questo deve essere tenuto in conto nella definizione dei valori di riferimento e dei criteri di valutazione;
2. allude anche alla 'qualità' dei dati. Questa viene generalmente associata alla proprietà del dato di rivelarsi adeguato (significativo) rispetto ad un certo fine. Mentre un errore dei dati può sempre essere espresso in termini assoluti, la qualità dei dati è funzione del contesto e dello scopo della loro rilevazione (una lista di criteri relativi alla qualità, include, ad esempio, l'accuratezza degli attributi, la coerenza logica, la completezza);

La nozione di 'incertezza' fa riferimento ad una qualche misura del 'dubbio' e della fiducia nei confronti di certi risultati. Si può essere certi dei risultati ottenuti a partire da dati poco accurati, di scarsa qualità e con errori. Viceversa, i risultati di una complessa serie di operazioni su dati accurati, precisi ed esatti, possono essere molto incerti.

[2] Secondo Casetti (1999):

Il modo di investigazione letterario si focalizza sulle descrizioni e sulle narrazioni. Per sua natura è flessibile e si rivela particolarmente adatto a trattare descrizioni di vario tipo provenienti anche da campi disciplinari diversi. E' particolarmente adatto a segnalare aspetti nuovi, o problematici di cambiamento della realtà

Il modo di investigazione accademico e' quello proprio che contraddistingue un campo disciplinare, che per sua natura possiede un quadro strutturato di conoscenze. Tale modo di investigazione e' l'esito di uno sforzo collettivo, volto a consolidare e sistematizzare in quadri concettuali, le spiegazioni che vengono sviluppate. Queste diventano oggetto di discussione e di riflessione e consentono l'istituzionalizzazione delle conoscenze entro un certo campo disciplinare.

Il modo di investigazione matematico e' fondato sull'uso di modelli matematici: Esso mette in relazione strutture analitiche 'formali' con concetti 'sostantivi' elaborati in uno degli altri due modi di investigazione. Una peculiarità di questo modo di investigazione risiede nel fatto che esso presuppone un 'controllo sul pensiero'

L'autore fa inoltre notare che per quanto si possa riconoscere un'evoluzione storica dei modi di investigazione (da un modo letterario, ad uno accademico ad uno matematico) e ciascuno di esso appartenga ad un campo che evolve, non ci sia rottura tra i tre modi di investigazione.

di molte cause, ciascuna con effetti modesti, allora i metodi statistici tradizionali o le reti neurali possono essere adatti (celle in alto a destra). Il ragionamento logico o simbolico (celle in basso a sinistra) è invece utilizzabile quando le cause sono poco numerose anche se gli effetti possono essere molteplici. Nessun metodo funziona in modo soddisfacente quando il problema presenta molte cause e molti effetti (a meno che il problema sia lineare e possa essere scomposto) (cella in basso a destra).Qualche volta è possibile trovare delle soluzioni accettabili cambiando il tipo di rappresentazione (ciò che semplifica la descrizione) o facendo convivere molteplici rappresentazioni.
Il più delle volte abbiamo di fronte problemi caratterizzati da molte cause ed effetti modesti. E' questo il campo dei programmi euristici e, più in generale, dell'intelligenza artificiale. In tale campo, i metodi analitici classici si rivelano spesso di scarso aiuto e devono essere integrati da metodi di ragionamento basati su analogie o casi di studio' (vedi anche celle centrali della Tab.2) (Minsky, 2000, p. 71).

Tabella 2 **Una tipologia di approcci ai fenomeni territoriali in relazione alla considerazione congiunta delle cause e degli effetti (elaborato sulla base di Minsky, 2000)**

	Numerosità delle cause		
Effetti	*Bassa*	*Discreta*	*Elevata*
Deboli	Interrogazione	Statistica lineare	Logica fuzzy, reti neurali
Modesti	Ragionamento qualitativo	IA classica	Ragionamento basato su analogie
Significativi	Ragionamento basato sulla logica simbolica	Ragionamento basato su casi	Problema intrattabile (????)

B) Con riferimento alla strumentazione tecnologica, lo sviluppo delle potenzialità di calcolo e, più in generale, la diffusione delle nuove tecnologie dell'informazione, stanno avendo un'influenza considerevole, non solo sull'architettura dei SIT, ma anche sulle metodologie per l'analisi urbana che a quell'architettura fanno riferimento. Fra le direzioni di sviluppo più promettenti vanno citate quelle volte a (Occelli, 1998):

- favorire l'integrazione dei due grandi filoni metodologici, relativi, rispettivamente, ai metodi (e alle tecniche) finalizzati(e) a descrivere ed investigare le componenti 'sostantive' degli oggetti di piano (in sostanza i metodi di analisi spaziale, la modellistica urbana e dei trasporti, i metodi input-output, ecc.); ed ai metodi (e le tecniche) orientati(e) al trattamento delle componenti 'procedurali' del processo di piano (in sostanza i metodi di valutazione e, più in generale, i modelli del processo decisionale).

- facilitare 'l'interfacciamento' sia tra procedure (package), finalizzati ad operazioni e compiti diversi (nella direzione ad esempio di relazionare in modo più efficiente dati, indicatori ed immagini visive) sia tra utilizzatori diversi (inter-operabilità fra ambienti di calcolo, comunicazione).

Una conseguenza di tale integrazione è testimoniata, ad esempio, dal recente revival di interesse per gli indicatori finalizzati alla valutazione sia delle performance urbane (Bertuglia, Clarke e Wilson, eds., 1994, Sawiki e Flynn, 1996) sia della sostenibilità, ambientale (Button, 2000)[3],[4].

C) Con riferimento, infine, agli utilizzatori, il nuovo modello di SIT si caratterizza per una crescente numerosità e diversificazione degli utilizzatori. Ciò presenta alcune implicazioni importanti, ancora largamente sottovalutate nella predisposizione delle nuove architettura dei SIT (Occelli, 1997):

- la prima è relativa ad un allargamento del tipo di utilizzatori/fruitori dei SIT. Non ci sono solo specialisti e/o esperti, che operano nel campo della progettazione e/o della pianificazione, ma anche professionisti ed utenti che lavorano in aziende private ed organizzazione pubbliche. Emerge in particolare una nuova figura di utilizzatore, quella del cittadino comune che, sempre più spesso è chiamato a partecipare attivamente al governo della propria comunità (Gyford, 1994);

- la seconda è che tale aumento di utenti porta con sé una maggiore variabilità delle percezioni e dei livelli di conoscenza circa i fenomeni territoriali ed i problemi di piano.

- La terza è che gli utenti stessi possono diventare al tempo stesso utilizzatori e produttori delle 'informazioni' messe a disposizione dai SIT[5].

Le trasformazioni delle componenti nel modello tradizionale di SIT prefigurano anche delle modificazioni nei requisiti attesi.

Una modificazione cruciale nel nuovo modello di SIT concerne, come già introdotto in precedenza, le aspettative conoscitive e, dunque, la natura stessa dell'informazione prodotta e/o che deve essere messa a disposizione. Se, nel modello tradizionale, il tipo di conoscenza richiesto doveva rispondere sostanzialmente ad una necessità generale di oggettivazione e di facilitazione procedurale, nel nuovo modello esso deriva da esigenze puntuali di costruzione di quelle condizioni locali di conoscenza che permettano di agire.

---

[3] Lo sviluppo degli indicatori urbani disaggregati a livello sub-urbano, in particolare, è alimentato da due principali ragioni (Sawiki e Flynn, 1996): a) la diffusione dei GIS, la cui tecnologia, sempre più sofisticata consente una rappresentazione delle fenomenologie urbane a diverse scale di risoluzione geografica e di dettaglio informativo; b) il progressivo spostamento delle responsabilità di governo (ad esempio nella fornitura dei servizi) dal livello centrale (statale) a quello periferico (locale) e l'enfasi posta sull'importanza delle partnership tra pubblico e privato. Ciò ha delle conseguenze significative in termini di esigenze conoscitive dei fenomeni urbani, le quali richiedono un'attenzione crescente alla loro articolazione territoriale ed alle loro specificità nelle diverse sub-aree.

[4] Si noti per inciso che questo stesso revival ripropone, sotto il profilo metodologico, un aspetto delicato della problematica degli indicatori, relativo alla scelta dell'unità e del livello di analisi, ed alla pertinenza di questi ultimi relativamente all'esigenza di disporre di indicatori 'utili' per la formulazione delle politiche urbane. A questo riguardo, riemerge, inoltre, il problema della significatività 'dell'articolazione territoriale' adottata rispetto alla quale predisporre gli indicatori. La soluzione ragionevole di ritenere che l'articolazione migliore possa essere quella che riflette le percezioni e/o i comportamenti dei residenti, si scontra tuttavia con un duplice ordine di problemi inerenti: a) l'intrinseca instabilità dei confini che, riflettendo le percezioni od i comportamenti dei residenti, possono essere mutevoli nel tempo; b) il riconoscimento della molteplicità dei contesti rispetto ai quali le azioni quotidianamente avvengono, ciò che si scontra con l'idea che il quartiere (l'unità di vicinato) possa essere il contesto più appropriato cui riferire lo studio dei comportamenti umani e delle azioni sociali.

[5] Quest'ipotesi stava alla base del Progetto di Osservatorio del Sistema Abitativo predisposto dall'Ires alla fine degli anni '80. Per ragioni diverse il progetto non venne realizzato, ma la sua concezione, basata appunto su un sistema di agenti fruitori/utilizzatori delle informazioni la cui interazione definisce i fabbisogni informative, rimane ancora valida.

Si tratta di aspetti che concorrono alla formazione di quello che alcuni autori hanno chiamato 'profound knowledge' (Anderson, Dooley e Misterek, 1991): di un'attività conoscitiva, cioè, che, per sua natura è continuativa, evolutiva e volta al miglioramento della qualità delle nostre azioni (miglioramento di processo) e degli esiti delle stesse (miglioramenti di prodotto).

Più in particolare, una tale conoscenza – che potrebbe essere denominata 'conoscenza incorporata' - è una conoscenza che (Innes, 1998)[6]:

- è capace di influire, poiché è incorporata nelle interpretazioni degli individui, nella pratica dell'agire corrente e nelle procedure istituzionali;
- essendo l'esito di un'interazione fra gli utilizzatori è portatrice di un significato condiviso;
- è costituita da una molteplicità di saperi.

Questo nuovo profilo della conoscenza coinvolge anche profondamente il ruolo che il metodo è chiamato a svolgere per la sua stessa produzione. A questo proposito, vi sono due aspetti specifici di modificazione del ruolo dei metodi che meritano di essere qui richiamati (Occelli, 1999):

- il primo è connesso all'estensione della funzione del metodo, da una funzione convenzionale – volta a garantire un'oggettivazione del percorso di analisi- verso una funzione che potremmo definire 'cognitiva', - consapevole dei propri modi di investigazione (Minsky, 2000)- finalizzata ad investigare o/ a scoprire rappresentazioni alternative di un certo problema;

- il secondo aspetto riguarda il processo 'di apprendimento', che sta alla base della costruzione di una conoscenza 'incorporata'. Poiché l'apprendimento si basa sul presupposto fondamentale di poter distinguere e valutare situazioni diverse al fine di migliorare le proprie possibilità di azione (Pednault, 2000), il metodo diventa un 'veicolo' fondamentale per operare tale distinzione e 'sperimentare' percorsi alternativi.

4. Considerazioni conclusive

Scopo di queste note era di evidenziare alcuni aspetti che contraddistinguono l'attuale evoluzione dei SIT. Esse hanno messo in luce come le modificazioni in corso coinvolgano sia la concezione generale sia le singole componenti dell'architettura dei SIT.

Un'indicazione generale che se trae è che le modificazioni osservate rispondono, in ultima istanza, ad un bisogno diffuso di accrescere la 'consapevolezza' circa le nostre possibilità di azione e gli esiti attesi e/auspicati di queste ultime sull'ambiente che ci circonda.

Dal punto di vista della concezione generale, le considerazioni avanzate indicano che il nuovo modello di SIT non prefigura in un'architettura specifica, ma propone invece una configurazione meta-organizzativa (che si ispira ai sistemi di tipo virtuale,Facheux, 1997) finalizzata a:

1. migliorare le capacità di osservazione e di esplicitazione degli aspetti non ovvii o 'taciti' (tramite in particolare le attività di sperimentazione);

[6] Un esempio di come tale idea di conoscenza incorporata si possa concretizzare è rappresentato dal lavoro contenuto in Spaziante (1999) sulle opportunità offerte dalle nuove tecnologie informatiche per l'integrazione funzionale dell'informazione territoriale tra diversi livelli decisionali.

2. 'concettualizzare il significato dell'azione' in modo da poterla riferire al (confrontare con il) contesto più generale, e, soprattutto, da poter rafforzare la capacità di risposta locale;
3. favorire la trasparenza delle azioni dei diversi attori rispetto al contesto.

In questa configurazione, inoltre, emerge come il ruolo del metodo acquisti una rilevanza crescente soprattutto per quanto riguarda la formulazione di descrizioni (i modelli e, più in generale, le rappresentazioni) che possano essere comprese e condivise da parte dei diversi utilizzatori del SIT.

Conseguentemente, anche il trattamento dei dati richiede di essere finalizzato ad una certa varietà di attività, quali:

- la performance dello stato di fatto (analisi delle operazione compiute e dei loro esiti)
- la valutazione critica (riflessione finalizzata a capire cosa si sta facendo);
- la comunicazione in ordine agli oggetti indagati;
- l'agire pratico, visto dal duplice punto di vista:dell'encoding (ovvero di quell'attività di astrazione che dà senso all'esperienza) e del decoding (ovvero di quell'attività attraverso la quale gli aspetti concettuali sono ricondotti nel contesto concreto);
- l'atteggiamento creativo, nella direzione di 'ricercare/esplorare' soluzioni inedite;
- l'affinamento continuo delle conoscenze apprese (consolidamento) e delle competenze.

Affinché il profilo meta-organizzativo che sta alla base del nuovo modello di SIT non configuri solo un'agenda ideale di requisiti desiderabili, occorre tuttavia creare le condizioni che ne rendano possibile la loro concreta applicazione nelle situazioni reali. Per quanto la loro individuazione non sia compito di queste note le seguenti indicazioni possono formularsi.

- Una prima condizione è rappresentata dal riconoscimento della complementarità delle attività conoscitive realizzate o sperimentate in un SIT. Nessuna singola attività, ristretta ad un- singolo campo disciplinare specifico (o limitata in un certo ambito istituzionale) è sufficiente per la creazione e la diffusione di 'conoscenza incorporata'. Esperienze diverse, anche realizzate in uno stesso ambiente, hanno un ruolo importante nel consentire di riconoscere situazioni problematiche o di individuare linee di azioni preferibili. E' ben noto peraltro, che le innovazioni più significative raramente hanno proceduto in modo lineare dai centri di ricerca o dalle università alle organizzazioni produttive di sviluppo dei prodotti. Teoria e pratica hanno spesso avanzato di pari in passo, in modo interattivo, coinvolgendo gli sforzi ed il supporto di enti e di organismi sia pubblici sia privati.

- Una seconda condizione è riferibili al fatto che la regola KISS (Keep It Simple Stupid) mal si adatta a gestire la complessità dei SIT moderni. L'introduzione delle nuove tecnologie di informazione, infatti, non solo muta la natura stessa dei problemi da affrontare (si pensi ad esempio al trattamento dei fenomeni di interazione di tipo virtuale,Janelle e Hodge, eds. 2000) ma influenza anche le nostre capacità di riconoscerli, di affrontarli e di gestirli. Ne consegue che il funzionamento della rete di utilizzatori in un SIT moderno dovrà misurarsi con esigenze spesso contrastanti, legate ad esigenze di flessibilità e di mantenimento di strutture preesistenti. Alcuni dei problemi fondamentali già segnalati al riguardo sono (Hasselbring, 2000): l'eterogeneità e l'autonomia dei singoli dei nodi locali, la distribuzione e la comunicazione fra i diversi nodi.

- Una terza condizione, strettamente connessa al percorso innovativo di cui sopra, è rappresentato dalla necessità della flessibilità. L'incertezza che accompagna le trasformazioni odierne spesso richiede variazioni e/o aggiustamenti negli approcci originariamente adottati. Ma può anche segnalare l'opportunità di proseguire un certo percorso metodologico anche se i risultati finali attesi non sono chiaramente definibili con precisione fin dall'inizio.

- Un'ultima ma non per questo meno importante condizione concerne l'opportunità di creare una 'comunità' di utilizzatori/agenti che sappiano cooperare per la costruzione e l'aggiornamento continuo dei SIT. Da questo punto di vista, la formazione dei diversi agenti, acquisita attraverso esperienze individuali in ambienti diversi, e la possibilità di spostarsi agevolmente fra tali ambienti (ad esempio tra università, imprese ed enti di governo) risulta una condizione importante da favorire.

Riferimenti bibliografici

Anderson J. C., Dooley K. J., Misterek S. D. A. (1991) The Role of Profound Knowledge in the Continual Improvement of Quality,*Human System Management, 10*, 143-259.

Andersson A. E., Batten D.F., Kobayashi K., Yoshikawa K. (eds.) (1993) *The Cosmo-Creative Society,* Springer, Berlin.

Bennett R.(1996) Administrative Systems and Economic Spaces, Relazione presentata al 28Th International Geography Conference, The Hague, 4-10 August.

Baherenberg G., Fischer M. , Nijkamp P. (eds) (1984) *recent Developments in Spatial data Analysis,* Gower, Aldershot.

Boulding K. E. (1985) The Sciences in the Spectrum of Human Knowledge,*Environment and Planning, 12,* 21-30.

Button K. (2000) City Management and Urban Indicators, Paper presented at the 6th World Congress of the Regional Science Association, Lugano, May 16-20.

Bertuglia C.S., Bianchi G., Mela A. (eds.) (1998) *The City and its Sciences*, Springer, Berlin.

Bertuglia C.S., Lombardo S., Occelli S. (1998) Nuove tecnologie dell'informazione e sistemi urbani:elementi di riflessione ed un'agenda propositiva, in Senn L, Boscacci F. eds., *I luoghi della trasformazione e dell'innovazione,* AISRE, SEAT, Torino.

Bertuglia C.S., Occelli S. (1995) Transportation, Communications and Patterns of Location, in Bertuglia C.S., Fischer M.M., Preto G. (eds.) *Technological Change, Economic Development and Space*, Springer, Berlin, 92-117.

Castells M. (1989) *The Informational City*, Blackwell, Cambridge.

Castells M. (1996) *The Rise of the Network Society,* Blackwell, Oxford.

Casti J. L. (1997) *Would-be Worlds*, John Wiley, New York..

Cecchini A. (1999) New Information and Communication Technologies and the Professions of Town and Regional Planners, *Computers Environment and Urban Systems, 23*, 151-172.

Couclelis H. (ed.) (1996) Spatial Technologies, Geographic Information, and the City, Technical Report, 96-10, NCGIA, Santa Barbara, CA.

Facheux C. (1997) How Virtual Organizing is Transforming Management Science,*Communications, 40*, 7, 50-55.

Fotheringham A.S., Brundson C. (1999) Local Form of Spatial Analysis,*Geographical Analysis, 31,4,* 340-358.

Graham S., Marvin S. (1996) *Telecomunications and the city*, Routledge, London.

Gyford J. (1994) *Citizens, Consumers and Council. Local Government and the Public*, Macmillan, London.

Hasselbring W. (2000) Information System Integration. Introduction,*Communications, 43*, 6, 32-38.

Innes J. E. (1998) Information in Communicative Planning,*Journal of the American Planning Association, 64,* 1, 52-63.

Janelle D. J., Hodge D.C. (2000) Information, Place, Cyberspace , and Accessibility, in Janelle D. G., Hodge D. C. (eds.), *Information, Place and Cyberspace*, Springer, Berlin, 3-11.

Janelle D. J., Hodge D.C.
(eds.), *Information, Place and Cyberspace*, Springer, Berlin.
Las Casas G., Storchi, G. (1994) Sistemi Informativi e analisi di grandi basi di dati, in Maciocco G. (a cura di) *La città, la mente, il piano. Sistemi intelligenti e pianificazione urbana,* Angeli, Milano, 419-435.
Lombardo S. (1994) Complessità, conoscenza e progettazione della città, in Maciocco G. (a cura di) *La città, la mente, il piano. Sistemi intelligenti e pianificazione urbana,* Angeli, Milano, 193-206.
Maciocco G. (a cura di)(1884) *La città, la mente, il piano. Sistemi intelligenti e pianificazione urbana,* Angeli, Milano, 193-206.
Maestre C. J. (1994) Histoire et contenu d'une approche des systèmes, *Analyse de Systèmes, 4.*
Nijkamp P., Rietveld P. (eds.) (1984) *Information Systems for Integrated Regional Planning*, North-Holland, Amsterdam.
Minski M. (2000) Commonsense-Based Interfaces, *Communications, 43,*8, 67-73.
Occelli S. (1997) Imprevedibilità e innovazione nella pianificazione : il ruolo della partecipazione, in Rabino G.(a cura di), *I volti del caso*, Angeli, Milano, 1997, 89-104.
Occelli S. (1998) A New Perspective for Methodologies in Spatial and Urban Analysis, in Bertuglia C.S., Bianchi G. and Mela A. (eds.) *The City and its Sciences*, Springer, Berlin, 851-872.
Occelli S. (1999) Perché è opportuno continuare a parlare di metodi. Lavoro preparato per il seminario, 'From Thinking to Action. Il ruolo delle metodologie nell'analisi dei sistemi socioeconomici. Problemi, esperienze e prospettive', Ires, 4/6/1999, Torino.
Occelli, S., Rabino, G.A (2000) Razionalità e creatività nella modellistica urbana, *Urbanistica, 113*, 22-26.
Pednault E. P. D. (2000) Representation is everything, *Communications, 43,* 8, 80-88.
Spaziante A. (1999) Dal 'mosaico dei PRG' strumenti per promuovere circolazione di informazioni e cooperazione fra livelli di governo del territorio, *Atti della XX Conferenza di Scienze Regionali*, Piacenza, 5-7 ottobre.
Unwin D. (1995) Geographical Information Systems and the Problem of 'Error and Uncertainty', *Progress in Human Geography, 19*, 549-558.
Turoff M. (1997) Virtuality, *Communications, 40*, 7, 38-43
Van Geenhuizen M., Nijlamp P. (1997) L'architettura dei sistemi informative nella gestione di una città complessa, in Bertuglia C.S., Vaio F. (a cura di) *La città e le sue scienze, vol.4, Le metodologie delle scienze della città,* Angeli, Milano, 1-66.

WORKSHOP

Sistemi informativi territoriali

Valorizzare il patrimonio informativo di SIT locali

A. Spaziante e P. Cipriano

Valorizzare il patrimonio informativo di SIT locali*

Agata Spaziante [1] e Piergiorgio Cipriano [2]

Dipartimento Interateneo Territorio del Politecnico di Torino
Viale Mattioli 39 - 10125 Torino
tel. 011.5647452 - fax 011.5647451 - e-mail: spaziante@archi.polito.it

Sommario

La tesi che s'intende dimostrare con questo contributo si riferisce alla possibilità ed utilità di integrare archivi e sistemi informativi territoriali locali al fine di valorizzare il patrimonio informativo delle amministrazioni comunali, disperso nei settori tecnici ed amministrativi, rendendolo efficace per la crescente domanda di conoscenze idonee al governo del territorio.

Piani e politiche urbane e territoriali hanno bisogno, infatti, di quantità e qualità crescenti d'informazioni per far fronte a processi evolutivi molto complessi e talora imprevedibili ed a dinamiche rapide e profonde.

Il patrimonio d'informazioni prodotte a livello locale per esigenze amministrative può essere molto utile a questo scopo, ma è sovente di difficile accesso e reperimento, a causa della sua dispersione nei vari rami delle amministrazioni comunali, della carente o assente documentazione, della mancanza di una ben definita strategia di gestione.

Connettere SIT comunali tra loro e con quelli livello superiore, renderli visibili all'esterno agevola un loro uso esteso e ne valorizza il significato a fini pianificatori, ma non solo. Soprattutto può promuovere la cooperazione fra Enti e fra servizi di uno stesso Ente e la diffusione e trasparenza delle informazioni cui ogni cittadino ha diritto.

A questo scopo è anche essenziale che tutte le informazioni amministrative comunali riferibili al territorio utilizzino un riferimento geografico comune (il codice toponomastico o quello della particella catastale) e siano documentate secondo formati standardizzati.

Avvalendosi di alcuni esempi tratti da ricerche del Dipartimento Interateneo Territorio del Politecnico di Torino con il CSI-Piemonte sulla Città di Torino, s'intende dimostrare la praticabilità e l'utilità a fini conoscitivi ed a supporto del governo del territorio, del contenuto di sistemi informativi territoriali comunali dotati delle caratteristiche corrette (georeferenziazione, documentazione) per una integrazione essenziale al loro uso multiplo ed esteso.

La tesi si avvale di un esempio relativo all'uso degli archivi gestionali del Comune di Torino connessi con l'attività edilizia per dimostrare la possibilità di produrre analisi utili a monitorare l'attuazione del PRG.

---

* Pur essendo il contributo risultato di un lavoro in collaborazione, il testo è stato steso da Agata Spaziante ed a lei fanno riferimento i paragrafi 1, 2, 3, 4. 7, 9, mentre a Piergiorgio Cipriano vanno attribuiti i paragrafi 5, 6, 8.

[1] Agata Spaziante è docente di Analisi della città e del territorio presso la I Facoltà di Architettura del Politecnico di Torino. E' responsabile scientifico del Laboratorio di Analisi e Rappresentazioni Territoriali ed Urbane (LARTU) del Dipartimento Interateneo Territorio. Coordina ricerche sull'uso delle informazioni per la pianificazione territoriale ed urbana e per la valutazione ambientale per Regione Piemonte, Provincia e Comune di Torino, CSI-Piemonte.

[2] Piergiorgio Cipriano è laureato in Architettura presso il Politecnico di Torino e titolare di una borsa di studio del CSI-Piemonte presso il Dipartimento Interateneo Territorio del Politecnico di Torino.

1. Introduzione

La questione alla quale questa relazione intende fornire un contributo riguarda il nodo di problemi che s'intreccia intorno all'uso dei sistemi informativi territoriali come strumento per il governo del territorio a scala locale ed intende sostenere la tesi che si debba realizzare una rete di sistemi informativi territoriali per valorizzarli superando la loro dispersione negli archivi gestionali dei settori tecnici ed amministrativi e rendendoli efficaci per la crescente domanda di conoscenze delle quali un settore particolarmente interessante è quello che fornisce supporto al governo del territorio.

Le riflessioni partono dagli aspetti più generali del problema per richiamare le numerose ragioni che confermano e rafforzano l'esigenza di rafforzare specificamente quelle caratteristiche che estendono la possibilità del loro uso al governo del territorio i sistemi informativi territoriali ed in particolare quelli che nascono dal patrimonio di dati gestionali prodotti e trattati dalle amministrazioni locali.

In situazioni di complessità e di incertezza, di vorticoso e rapido cambiamento quali sono quelle che i nostri territori attraversano oggi, diventa sempre più rilevante il significato che assume la fase della conoscenza nel processo di pianificazione. E dunque cresce il ruolo che le questioni relative all'informazione hanno per chi di pianificazione si occupa, crescono le attese nei confronti del contenuto che oggi ci si attende di trovare nei sistemi informativi, soprattutto si conta quasi in una loro miracolosa capacità di aiutare a svelare le cause ed a prevedere gli effetti dei processi che le scienze del territorio sempre più faticosamente registrano ed interpretano.

Cresce anche la domanda di uso dei sistemi informativi al di fuori delle finalità consuete: per la protezione civile, per l'organizzazione della logistica, per viaggiare. Si aprono quindi infinite possibilità di mettere a frutto questi strumenti.

Per contro a queste esagerate attese si contrappone l'ancora modesta diffusione di sistemi informativi tecnologicamente avanzati presso le amministrazioni locali e, soprattutto, anche là dove l'impianto di tali sistemi è avvenuto, va lamentato lo scarso uso (in termini sia quantitativi sia qualitativi) che di essi si fa nei processi di pianificazione, di attuazione dei piani, di amministrazione ordinaria del territorio, di offerta di servizi e di informazioni al cittadino, all'operatore privato, al gestore di funzioni pubbliche.

La tesi che ci proponiamo di sostenere si può riassumere nell'affermazione che attraverso la valorizzazione del patrimonio informativo gestito dalle amministrazioni locali e soprattutto attraverso l'integrazione di questi ultimi con altri sistemi informativi di diverso livello e di diverso gestore, è possibile conseguire un valore informativo aggiunto che può accrescere il significato e gli attributi dei singoli contenuti e incrementarne la qualità e l'uso tanto da contribuire a superare alcuni dei limiti da cui dipende la modesta efficacia delle conoscenze per il governo del territorio ed in qualche misura la stessa ridotta efficacia della pianificazione.

Ciò va incontro anche alle necessità che oggi la società civile, viziata dall'uso amplissimo della rete Internet per le funzioni più varie, pone con forza crescente all'Ente pubblico dal quale si aspetta di ottenere trasparenza e facilità di accesso alle informazioni.

In questa direzione dunque occorre concentrare le sperimentazioni e le iniziative di enti ed istituzioni che detengono grandi patrimoni informativi ed a questo obiettivo dovrebbe mirare anche l'ISTAT, che controlla una delle più ricche e consolidate fonti di informazioni e può svolgere un importante ruolo di stimolo nei confronti di altri enti gestori di informazioni, in particolare di quelli a livello nazionale, affinché si trovino le soluzioni per integrare i sistemi informativi.

In una prima parte si esaminano dunque le premesse generali dalla questione: la criticità e la responsabilità elevate che assumono gli strumenti della conoscenza quando il loro fine è quello di sostenere l'azione del governare il territorio; l'alto livello di complessità che caratterizza le conoscenze sui sistemi urbani; la crescente importanza che le nuove leggi urbanistiche e le nuove forme del piano (quelle che intendono distinguere una fase strategica da una fase operativa) attribuiscono ai momenti della descrizione e dell'interpretazione ed agli strumenti funzionali ad essi; i contenuti nuovi e complessi che i sistemi informativi devono proporre, e più ancora predisporre per il futuro, per rispondere adeguatamente alle esigenze di chi guida le trasformazioni i sistemi territoriali che stanno mutando dimensione, contenuto, forma, dinamiche, criticità.

In una seconda parte, prendendo spunto da alcune ricerche condotte dal Dipartimento Interateneo Territorio del Politecnico di Torino con ilCSI-Piemonte ed altri partners (il Comune di Torino, la Regione Piemonte) la tesi viene sviluppata attraverso una rassegna via Internet dei capoluoghi di Provincia dotati di Sistemi Informativi Territoriali per valutare la diffusione di strumenti innovativi di trattamento delle informazioni gestionali dei Comuni ed attraverso una verifica condotta su alcuni archivi gestionali del Comune di Torino, attorno al problema dell'analisi dell'attività edilizia locale e del monitoraggio dell'attuazione del nuovo PRG della città.

Le conclusioni intendono raccomandare alle pubbliche amministrazioni un'azione concentrata soprattutto su alcune condizioni essenziali perché la valorizzazione di questi archivi diventi possibile.

2. Conoscere per governare la città e il territorio

Occuparsi di sistemi informativi territoriali può significare affrontare la questione delle informazioni sotto molti punti di vista.

Quello che qui viene privilegiato, anche in considerazione dell'ambito in cui la sperimentazione si è svolta[3], è il punto di vista di chi utilizza i sistemi informativi ma nello stesso tempo contribuisce a definirne contenuti, formati, interazioni, elaborati, trattamenti ponendosi dalla parte di chi fa o gestisce piani, ed in particolare piani comunali. Dunque intendiamo discutere di sistemi informativi soprattutto dalla parte degli utenti particolari quali sono architetti e ingegneri che fanno ed usano i piani; amministratori che li discutono e li approvano; consulenti che ai piani e con i piani lavorano; operatori economici che con i piani devono fare i conti; cittadini che dovrebbero volerli capire e condividere o rifiutare.

I piani ed il governo del territorio da sempre si misurano con l'apparato conoscitivo disponibile e certamente incidono sulle sue caratteristiche, sebbene forse in misura inferiore a quanto sarebbe legittimo aspettarsi, e viceversa caratteri e contenuto della conoscenza condizionano (anch'esse in termini forse insufficienti) le possibilità e modalità di governare il territorio. E' chiara comunque l'importanza del ruolo che la conoscenza è chiamata a svolgere in questo contesto: per fornire supporto a decisioni rilevanti come quelle che comportano trasformazioni del territorio non basta descrivere e rappresentare ma occorre almeno interpretare e più ancora intuire tendenze, valutare effetti, proporre correzioni di rotta a chi professionalmente

[3] Questo contributo si avvale dei risultati di una ricerca condotta da Agata Spaziante e Carlo Alberto Barbieri del Dipartimento Interateneo Territorio del Politecnico di Torino per ilCSI-Piemonte (Consorzio per il sistema informativo di Regione Piemonte, Università e Politecnico di Torino) e dell'attività di ricerca condotta da Piergiorgio Cipriano con una borsa di studio delCSI-Piemonte presso lo stessa Dipartimento Interateneo Territorio.

è chiamato a produrre progetti e piani (e dunque ad apportare, attraverso questi, modifiche importanti a spazi e funzioni), sostenendo con l'apporto tecnico-scientifico l'azione politica di orientamento verso obiettivi di interesse collettivo. E dunque si tratta di cosa ben diversa dalla disponibilità di dati che un qualunque Sistema Informativo può accumulare ed archiviare. L'apparato conoscitivo di cui la pianificazione ha bisogno è uno strumento caricato di forti responsabilità e di particolari difficoltà, perché ha comunque a che fare con sistemi complessi, sia si tratti di quelle agglomerazioni urbane in cui le funzioni si addensano producendo degrado fisico, emarginazione, criminalità, inquinamento, congestione, sia che si tratti, al contrario, di aree in cui degrado, emarginazione e criminalità sono l'effetto della rarefazione, dell'indebolimento, dell'impoverimento economico e sociale delle comunità.

Dunque chi ha a che fare con i piani usa dati e informazioni ma li trasforma in interpretazioni, previsioni, norme, controlli, valutazioni, bilanci. E dunque esprime esigenze informative sofisticate, mirate, diversificate e mutevoli nel tempo e nello spazio.

Per queste ragioni, com'è noto, il rapporto analisi - piano non è sempre stato facile: ci sono stati momenti in cui ai limiti degli strumenti metodologici di analisi ed alla carenza di informazioni si sono attribuite rilevanti responsabilità negli insuccessi della pianificazione, nei risultati sconfortanti dell'urbanizzazione selvaggia, nei danni irreversibili prodotti dall'uomo all'ambiente. La mancanza di informazioni è stata indicata come una delle ragioni importanti della crisi delle analisi davanti all'inestricabile intreccio delle contraddizioni urbane ed una causa importante dell'incapacità delle scienze territoriali a fornire spiegazioni e previsioni, dell'urbanistica a suggerire strumenti di correzione e di orientamento delle trasformazioni, della politica a proporre obiettivi responsabili e condivisi.

Oggi, in una fase più matura di queste riflessioni ed in un momento più solido del dibattito disciplinare sul ruolo, sulle caratteristiche, sul trattamento e sulla diffusione delle informazioni per la pianificazione, siamo tutti consapevoli che la quantità di informazioni non è mai stata in realtà insufficiente a questi fini, che è stata piuttosto la qualità a rivelarsi insoddisfacente e che la sfiducia nelle capacità della tecnica di produrre conoscenze adeguate è stata sovente l'alibi e la copertura per mascherare una crisi della pianificazione di origini ben più profonde e lontane ed ha enormemente influenzato le pratiche di governo del territorio. Ci appare pertanto meno accettabile, oggi, la rinuncia a pianificare città e territori con strumenti generali e la linea del trasformare la città per parti e per progetti, giustificata anche con l'impossibilità di disporre delle conoscenze idonee a governare il sistema nel suo complesso.

Stiamo dunque attraversando una fase in cui questo ruolo (perché e per chi sapere) sta crescendo per diversi motivi:

- si sono andate diffondendo negli ultimi anni proposte ed ipotesi a favore di un ruolo molto forte delle conoscenze nel processo di pianificazione, per produrre descrizioni che fungano da termine di riferimento per ogni proposta di intervento nonché per la definizione dei parametri e dei criteri di valutazione degli interventi stessi, per garantire il monitoraggio dell'attuazione del piano. Ne costituisce conferma il risalto che le nuove Leggi Regionali sull'uso del suolo varate negli ultimi anni (quelle della Regione Toscana, Emilia Romagna e Liguria, in particolare) attribuiscono ai momenti conoscitivi dell'elaborazione ed approvazione del piano[4]

[4] Si possono qui ricordare, ad esempio, gli specifici articoli che la recentissima Legge 217/2000 "Disciplina generale sulla tutela e l'uso del territorio" dell'Emilia Romagna dedica alla questione delle conoscenze nel processo di pianificazione del territorio:art. 17 "Coordinamento e integrazione delle informazioni" che promuove l'integrazione e implementazione del quadro conoscitivo del territorio attraverso il concorso e l'intesa fra Regione

- in queste nuove forme del piano (piani strutturali e piani operativi), che attribuiscono dunque una funzione molto significativa e non solo rituale alla descrizione ed all'interpretazione (anche al fine di individuare gli elementi strutturali su cui appoggiare le strategie) occorre a maggior ragione pensare non solo a nuovi contenuti ed a nuovi modi di produzione delle conoscenze, ma soprattutto ad un loro nuovo uso ed alla necessità di vederle alimentate da un flusso continuo e bidirezionale di informazioni fra i soggetti più direttamente interessati (Comuni, Provincie, Regioni), anche per consentire bilanci e monitoraggi della pianificazione che diventano ancor più fondamentali là dove si utilizzano le nuove forme concertate di governo del territorio al di fuori dal piano (conferenze dei servizi, accordi di programma, programmi integrati, ecc.)
- i problemi della complessità a cui in realtà l'urbanistica e le scienze del territorio, pur attraverso dibattiti e crisi della stessa idea di piano, non si sono mai sottratti, certamente oggi sono resi ancor più difficili dalla consapevolezza che oltre a dover assicurare le opportunità di un equilibrato sviluppo del territorio il governo del territorio deve perseguire la ricerca di condizioni che rendano queste evoluzioni "sostenibili" per l'ambiente naturale ed artificiale molto provati dalle modalità irresponsabili con cui queste sono avvenute ciò comporta integrare nel piano nuovi aspetti e nuove discipline che affrontino questi nuovi problemi.

L'insieme di questi aspetti lancia una nuova e più difficile sfida a metodi e tecniche della conoscenza e rende il ricorso a sistemi informativi sempre più urgente e necessario. Richiamano infatti la presenza di dinamiche differenti, non riducibili a correlazioni semplici e lineari, e dunque la possibilità che ad ogni azione di un attore strategico corrisponda una risposta fortemente condizionata dagli altri attori complementari, la necessità di introdurre aspetti valutativi sulla compatibilità delle soluzioni possibili con obiettivi di ordine superiore come sono quelli legati alla tutela delle risorse primarie, l'attenzione alla dimensione tempo, che comporta una diacronici delle informazioni.

L'evoluzione delle città e del territorio si muove dunque entro la complessità e la contraddittorietà di queste azioni incrociate e di questi obiettivi conflittuali e pone pertanto l'esigenza di contemplare anche l'imprevedibilità delle azioni causali ed il caos come dinamiche possibili: compiti ardui per chi governa il territorio e non meno ardui per chi deve proporre al decisore gli elementi per capire e scegliere.

3. Nuove domande di conoscenze per un territorio in evoluzione

---

ed enti locali; art. 51 "Monitoraggio e bilancio della pianificazione" che definisce compiti diversi e complementari per Provincie e Regione nella raccolta e gestione degli archivi della strumentazione urbanistica per redigere rapporti periodici sullo stato di attuazione dei piani; Capo A-VI "Strumenti a supporto della pianificazione - Strumenti cartografici" dedicato a definire le specifiche della rappresentazione cartografica digitale di tutta la cartografia e delle informazioni relative al territorio ed alla pianificazione a scala regionale.
In termini non meno precisi e rilevanti le Leggi regionali di Toscana e Liguria sanciscono ruoli e formati dell'apparato conoscitivo di supporto alla pianificazione.
Anche la proposta della Giunta Regionale del Piemonte per la nuova Legge urbanistica regionale, sia pure molto più debolmente, all'art. 3 su "Mezzi di conoscenza e informazione" dà mandato alla Regione perché provveda in collaborazione con Provincie e Comunità Montane a raccogliere i dati per la conoscenza continua del territorio utilizzando le tecnologie più idonee, anche di natura informatica, ed utilizzando tecniche idonee per conseguire con efficacia e rapidità le finalità della Legge.

Anche i contenuti della conoscenza (su cosa dobbiamo sapere) vanno innovati:

- gli elementi oggi necessari per pianificare si arricchiscono di temi nuovi (basti ricordare le tematiche ambientali, la qualità insediativa, la conservazione o il recupero di identità locali), ma anche di referenti nuovi (nuovi livelli di pianificazione, come quello provinciale, nuovi soggetti come le società miste e le Agenzie) e dunque vanno diversificate in relazione alle sedi istituzionali ma non frammentate, per quanto riguarda responsabilità di produzione, piuttosto che possibilità di utilizzo delle stesse. Un esempio: le tematiche ambientali a scala vasta richiedono quadri conoscitivi costruiti sulle specificità degli ambienti locali, ma anche capacità di controllo ed indirizzo strategico di area vasta (le competenze in materia ambientale sono oggi attribuite soprattutto a Regione e province).Allo stesso tempo politiche ambientali di scala comunale presuppongono - per essere efficaci- un confronto con descrizioni condivise a vasta scala, che solo le province o le Regioni sono in grado di fornire.
  La possibilità, quindi, per chi governa il territorio in termini di area vasta, di arrivare a costruire tematismi utili, a partire dalla pianificazione comunale, costituisce oggi un obiettivo irrinunciabile, in vista anche del potenziamento di un sistema informativo (territoriale, ma anche collegato alle decisioni informali, di "volontà e desideri", che costituiscono di fatto un valido supporto alle future decisioni di Piano) tra i diversi livelli amministrativi.
  La domanda di conoscenze su base territoriale utili ad affrontare questa serie impegnativa di questioni per governare il territorio è dunque in continua crescita.
  Il territorio è indagato da decenni alla ricerca di risposte ai quesiti sui fenomeni che fanno nascere, crescere, declinare, risorgere le aggregazioni umane secondo percorsi che ci sfuggono, ed ancor più nella speranza di trovare metodi convincenti per guidare le trasformazioni verso obiettivi prefissati.
  Alla insoddisfazione per la modestia delle risposte che le scienze del territorio hanno saputo dare finora a queste domande, si aggiunge oggi la consapevolezza che, parallelamente ad un gigantesco progresso delle tecnologie del sapere, è avvenuto un altrettanto gigantesco cambiamento dell'oggetto delle nostre investigazioni perché il territorio e soprattutto la città oltre che più dinamici sono per certi aspetti divenuti molto diversi da quelli che ci eravamo abituati ad esaminare nei decenni precedenti.

- L'affermarsi, nel contesto dei Paesi ad economia sviluppata di una logica della trasformazione senza crescita, dello sviluppo orientato alla qualità ed al miglioramento delle condizioni di vita della popolazione in sostituzione di una logica dell'espansione, ha spostato in secondo piano i fenomeni quantitativamente manifesti e più dotati di informazioni sistematiche, che costituivano l'oggetto privilegiato di attenzione delle analisi urbane e territoriali (quantità e densità di aree destinate alle diverse funzioni, quantità e tipo di popolazione e di attività economiche ecc.) e privilegiato fenomeni percepibili soprattutto in termini qualitativi, rilevabili attraverso segnali più sofisticati e sottili e meno dotati di conoscenze tradizionali e sistematiche (l'evoluzione nell'uso degli immobili, le caratteristiche sociali della popolazione, il contenuto professionale del lavoro, il livello delle funzioni economiche): certamente una sfida per il patrimonio conoscitivo necessario a capire e seguire il cambiamento.

- Inoltre è in corso una vera e propria sfida tra le città sul terreno della loro capacità di conciliare due contrapposte esigenze: lo sviluppo di innovazione (necessaria a farle divenire o rimanere i principali luoghi della nuova politica culturale e delle attività del comando, della decisione, dell'invenzione) e nello stesso tempo la ricerca di "sostenibilità" ambientale (necessaria a garantire la riserva di risorse per le generazioni future). Anche per mirare a questo difficile equilibrio e dunque per migliorare la propria organizzazione interna, le città ed i loro amministratori devono contare, come fondamentale strumento di azione, sull'efficacia delle loro conoscenze.

Per guidare le trasformazioni è dunque necessario interpretare e seguire vecchi e nuovi aspetti della realtà urbana: i cambiamenti nelle dinamiche demografiche ed i problemi dell'integrazione con il popolo degli immigrati, la domanda di spazi per nuove funzioni o di luoghi nuovi per funzioni tradizionali, la crescita della città in forme illegali, l'aumento del benessere e della domanda di consumi e l'uso sconsiderato delle risorse (acqua, aria, suolo) ma anche lo smaltimento incontrollato e irresponsabile dei rifiuti, l'incentivo ad un modello di urbanizzazione diffuso e la progressiva sottrazione dei suoli ad elevata fertilità all'uso agricolo, l'aumento di pendolarità e di inquinamento, la rivoluzione nella struttura del mercato del lavoro e l'aumento della "povertà senza assistenza", la piaga dell'emarginazione e della violenza ed il bisogno di "sicurezza" individuale e di "protezione civile", la ricerca di identità e di qualità per i luoghi delle periferie e l'imperativo della salvaguardia e della valorizzazione del patrimonio storico dei centri urbani.

Di fronte a questo impegnativo compito gli studi urbani sono in sempre più grave ritardo nello sviluppo di metodi e tecniche idonei a superare il livello della comprensione per proporre strumenti di previsione, come nei passati decenni, durante le fasi più alte e più ottimiste delle scienze regionali, quando ci si è illusi di poter mettere a punto modelli capaci di controllare le variabili strategiche dello sviluppo.

Oggi, in un momento di minore fiducia e di maggiore riflessione, il terreno su cui si concentrano gli studi non è più quello degli strumenti per prevedere, quanto quello delle tecniche per studiare, capire e razionalizzare la relazione causa – effetto, per spiegare la relatività dell'interpretazione e della realtà stessa. La formalizzazione, dunque, se non fornirà previsioni, fornirà un modo chiaro attraverso cui confermare la complessità e la significatività relativa dello stesso mondo reale ed osservare in modo continuo le trasformazioni per misurare l'efficacia delle azioni di governo intraprese.

Occorre però prepararsi ad affrontare altre e più nuove forme della urbanizzazione e della trasformazione delle città, oggi ancora imprevedibili. Già oggi stanno entrando tra le componenti oggetto di attenzione e stanno portando a profonde revisioni negli strumenti di governo del territorio molte nuove analisi destinate a sondare aspetti sfuggiti fino ad ora agli studi urbani tradizionali quali gli effetti indotti dall'inquinamento e dalla dissipazione dei flussi di risorse; la ricerca di efficienza economica e di equità sociale; la valorizzazione delle identità locali e dei valori comunitari Certamente possiamo intuire che altri aspetti ancora poco praticati si stanno rivelando strategici e porranno nuove domande di conoscenza di sempre più difficile contenuto, quali l'attenzione al mantenimento dei cicli naturali ed alla capacità di sopportazione delle pressioni sull'ambiente nonché alla irreversibilità dei processi energetici, gli studi sulla stabilità dei sistemi, il ruolo della dimensione*tempo*.

Queste nuove esigenze degli studi e degli strumenti di governo del territorio spingeranno le amministrazioni locali a dotarsi di strumenti più sofisticati di conoscenza, in termini di tecnologie ma soprattutto in termini di contenuti informativi

3. Le parole-chiave del sistema delle conoscenze per il governo del territorio

Oggi dunque all'apparato delle conoscenze si chiede molto.

Soprattutto si chiede di mettere i soggetti del governo del territorio in grado di affrontare, a qualunque scala di governo del territorio, i connotati nuovi che piani e politiche, tutti, devono assumere per adattarsi alla mutata natura delle trasformazioni.

Qui se ne sottolineano alcuni tra i più importanti e nuovi, che costituiscono vere e proprie *parole-chiave* su cui costruire le priorità della ricerca e della pratica in materia di conoscenze necessarie al governo del territorio:

- *la processualità*, nei rapporti verticali tra livelli istituzionali, nei rapporti orizzontali tra diversi attori pubblici e sociali, nella successione dinamica tra i diversi piani e tra le successive fasi di attuazione;
- *la qualità dell'organizzazione territoriale* che assicuri condizioni funzionali articolate in grado di soddisfare le esigenze tanto dei singoli operatori quanto delle funzioni territoriali indivisibili;
- *l'equità distributiva* di costi e benefici tra i diversi soggetti e le diverse parti del territorio, agendo sui molti strumenti legislativi, finanziari, tecnico-urbanistici, da quelli più tradizionali (come gli oneri di urbanizzazione) a quelli più innovativi (come i meccanismi perequativi), a disposizione delle istituzioni locali attraverso o insieme ai piani per conseguire un risultato meno squilibrato di quello che abitualmente si produce attraverso i processi di governo del territorio;
- *la trasparenza*, come obiettivo di legittimazione ottenuto attraverso la formazione di un consenso sociale allargato ;
- *la fattibilità*, come ricerca di compromesso fra rigidità nei confronti di indirizzi strutturali e non negoziabili, necessariamente prescrittivi, e flessibilità nei confronti delle azioni programmate di attuazione di tali indirizzi, inevitabilmente concertate.

Il ruolo del sistema delle conoscenze nei confronti di esigenze così vaste e complesse viene sicuramente rafforzato e dilatato fino a diventare per alcuni versi una delle criticità per il passaggio ad una nuova fase della vicenda urbanistica.

Soprattutto diventano essenziali alcuni attributi che rendano la natura ed i contenuti delle conoscenze coerenti con la visione nuova verso cui parrebbe muoversi il sistema degli strumenti di governo del territorio:

- *continuità*, per essere in sintonia con un piano che non è più la definizione di un assetto, con un controllo, che non è più il mero riscontro di conformità una tantum delle destinazioni d'uso rispetto alla normativa, con una gestione che non si attua più mediante il confronto dei progetti attuativi con le determinazioni del piano, con un'azione di governo del territorio che si esplica attraverso una sequenza continua di momenti decisionali caratterizzati da forza, durata, contrattabilità, interlocutori differenti ;
- *flessibilità*, per affrontare l'incertezza dei processi sociali ed economici in un contesto sempre più dominato da elementi esogeni che la globalizzazione dei mercati importa e

sovrappone, quasi ovunque, sulle scelte locali, ma anche per adattare le azioni locali ad una gamma di soggetti sempre più varia, imprevedibile, mutevole

- *interattività*, con i diversi livelli istituzionali ed i diversi soggetti e destinatari del piano, da quelli collettivi ed organizzati fino a quelli privati ed individuali, per garantire l'accessibilità alle scelte ed ai contenuti, per confrontare e diffondere informazioni, dati, interpretazioni e rappresentazioni, al fine di ottenere un consenso sociale largo intorno all'azione di governo.

Con quali mezzi, allora, rispondere a così ambiziose esigenze da cui dipende la reale possibilità di comprendere e prevedere il comportamento dei sistemi urbani?

Come utilizzare le tante informazioni che popolazioni ed attività depositano in tanti diversi contenitori di dati e come accedere alle tecnologie, oggi potentissime e non più molto costose e difficili, da cui possiamo farci aiutare per trarre da questo materiale informe significati e prospettive utili a guidarne i comportamenti verso obiettivi predefiniti e concordati democraticamente, come il nostro sistema istituzionale ci richiede?

3. Dai SIT informazioni per chi governa il territorio... ma non solo

Il contributo che i SIT possono dare ad una domanda di conoscenza così complessa è ben noto. La loro stessa definizione come "*sistemi informatizzati per l'acquisizione, la memorizzazione, il controllo, l'integrazione, l'elaborazione e la rappresentazione di dati spazialmente riferiti alla superficie terrestre*"[5] e l'accento che le parole-chiave scelte per indicarli pongono sull'integrazione delle componenti (donde il termine *sistema*), sull'aspetto interpretativo associato al termine *informativo* e sulla specificità dei dati *territoriali* trattati, contengono gli elementi essenziali del ciclo completo di un processo informativo costruito per rispondere adeguatamente alle molteplici e complesse esigenze elencate nei precedenti paragrafi.

Dunque il percorso verso una risposta adeguata al nuovo ruolo ed ai nuovi contenuti che la conoscenza è chiamata ad offrire a chi di governo del territorio si occupa, sotto qualunque punto di vista, passa inevitabilmente attraverso l'innovazione nel modo in cui i dati sono trattati e quindi attraverso la realizzazione di SIT.

Oggi questi sistemi sono in grado di rappresentare, per le Pubbliche Amministrazioni competenti in materia di politiche territoriali, uno strumento efficace a questo fine e ciononostante la loro diffusione ed il loro uso nel nostro Paese non è ancora soddisfacente ed è anzi ampiamente sottodimensionato a confronto di quanto avviene in altri Paesi europei in situazione economica, sociale ed urbanistica non dissimile dalla nostra.

I motivi che ne ritardano l'adozione presso gli Enti territoriali alle diverse scale o che ne limitano l'uso, sono molti. Più che alla tecnologia, ormai potentissima non più così costosa come era stata nelle fasi iniziali della loro introduzione ed ormai ben retta dall'ampia disponibilità di competenze informatiche, queste difficoltà vanno ascritte soprattutto a:

- indisponibilità o inadeguatezza dei dati (basi cartografiche digitali e datialfa-numerici georiferiti)
- carenze negli elementi organizzativi (ovvero insufficiente coordinamento delle risorse tecniche, delle persone, dei loro compiti, delle competenze degli Enti)

[5] Arnaud A., Masser I., Salge F., Scholten H., 1993, "GISDATA Research Programme", *European Science Foundation GISDAT NewsLetter,* n. 1

- insufficienza della formazione delle persone (intese sia come tecnici non informatici in grado di gestire l'alimentazione e l'uso del sistema, sia come utenti ovvero urbanisti ed esperti delle varie discipline che trattano dati territoriali e li trasformano in informazioni mirate, ma anche operatori privati e cittadini)

Fra dati, organizzazione, persone e tecnologie, l'elemento forte è costituito dalle tecnologie, che non rappresentano più un problema e forse non l'hanno rappresentato mai[6], mentre dati, organizzazione e persone vanno considerati altrettanti punti deboli: l'introduzione di un SIT là dove i problemi relativi a questi ultimi aspetti non siano stati risolti non solo rischia di produrre pochi benefici in termini di efficienza della struttura, ma addirittura potrebbe esaltare le criticità delle componenti assenti o non portate ad un livello quantitativo e qualitativo adeguato.

Si potrebbe affermare pertanto che mentre l'efficienza di un SIT dipende dalla qualità della sua tecnologia, la sua efficacia dipende da quella dei dati, delle competenze e della struttura organizzativa in cui viene inserito.

Tra le molte criticità citate se ne vuole qui riprendere in particolare una, quella relativa alla *indisponibilità dei dati*, per ribadire che in realtà il problema non è costituito dalla loro scarsa quantità: sono infatti copiosi i dati *georeferenziati o georiferibili*[7] che molti Enti, sovente sovrapponendosi più o meno consapevolmente, producono e trattano per loro fini istituzionali e utilizzano solo in piccola parte. E' più che giustificata, in questo senso, l'accusa alla Pubblica Amministrazione di sprecare informazioni ma anche risorse e di contribuire anche per questa via all'inefficienza del sistema.

Il più grave problema è in realtà costituito dalla ridotta accessibilità e circolazione del dato.

A questo proposito si devono aggiungere almeno altre tre molto importanti considerazioni:

- la disponibilità di informazioni su basi territoriali strutturate nei SIT e le opportunità offerte dalla tecnologia delle telecomunicazioni rendono molti dei dati geografici prodotti e gestiti da un'Amministrazione o da una sua parte utili e talora essenziali anche per altre Amministrazioni o per altre parti di essa, tanto da poter essere introdotte nel loro SIT secondo punti di vista e competenze del tutto diversi (ne è un esempio l'uso a fini tributari comunali degli archivi catastali o l'uso a fini di analisi sulle dinamiche delle attività economiche o di controllo dell'abusivismo edilizio di dati tratti dal catasto elettrico). Diventano così elementi strategici due altre dimensioni del SIT: l'*interscambiabilità* dei dati fra diversi utenti e la *pubblicizzazione* del loro contenuto informativo. Occorre rendere noto quel che le varie amministrazioni producono ed archiviano e quel che è possibile condividere. Ciò richiama la centralità di due aspetti relativi al problema della disponibilità dei dati, che verranno più ampiamente ripresi in

[6] La tecnologia GIS ha offerto sempre prestazioni molto più sofisticate di quelle più diffusamente applicate.

[7] "La georeferenziazione è prendere un indirizzo, e renderlo visibile come un "punto" posizionato su una mappa geografica o topografica. Ossia trasformare un indirizzo da entità astratta e puramente concettuale in un elemento concreto e immediatamente percepibile quale è un punto sulla carta topografia di una città o di un continente, o anche di un oceano, per renderlo riferibile - e quindi collegabile - ad altri punti più noti, come ad esempio il proprio indirizzo, trasformato anch'esso in un puntogeoreferenziato.
Ma se la georeferenziazione servisse solo a questo, avremmo creato un mastodontico giochetto informatico solo per multimiliardari: ne abbiamo invece fatto unsemplificatissimo strumento di lavoro […]."
(Giorgio Artuffo, in MondoGIS - Rassegna Stampa, http://www.domino-research.it/rass_stampa.htm)

seguito: la loro *documentazione* (quali, prodotti come, residenti dove, ecc.) e la loro *accessibilità* (chi può disporne, come, quando, a quali condizioni, ecc.)

- le potenzialità aperte dall'avvento delle tecnologie delle reti, e di Internet in particolare, che hanno reso l'accesso ai dati in generale facile e poco costoso, sta favorendo la domanda di informazioni basate su questi sistemi ma destinate a molteplici utenti anche non tecnici (dagli operatori economici ai cittadini) e per usi i più vari (dal turismo alle funzioni di consumo, alla gestione delle emergenze, alla partecipazione sociale) attraverso la sovrapposizione alla struttura di base (la cartografia digitale ed un corpo essenziale di dati georeferenziati su popolazione, attività economiche, usi del suolo) di livelli informativi i più disparati trattati da intermediari che li arricchiscono e finalizzano ai più diversi scopi

- la diffusione rapida, al di fuori dei loro campi di applicazione tradizionale, dell'uso di informazioni georeferenziate in formato digitale sta producendo da una parte una situazione di mercato caratterizzata dalla forte richiesta di questo genere di informazioni sollecitato dalle ditte di software specifico (si pensi alla rapida diffusione dei sistemi di assistenza alla guida degli autoveicoli, di controllo delle reti logistiche, della sicurezza di elementi in movimento, ecc. da parte di sistemi esperti appoggiati a SIT), dall'altra sta inducendo le pubbliche Amministrazioni a mettere il loro patrimonio informativo a disposizione non solo degli altri Enti Pubblici ma anche dell'utenza vastissima che consulta la rete Internet, immettendo parti consistenti dei propri SIT sui propri siti *Web* e venendo così incontro alle giuste attese di trasparenza e di partecipazione che le nuove normative sul rapporto fra cittadini e Pubbliche Amministrazioni hanno creato e sovente non assecondato con alcuna azione. Oggi attraverso la rete Internet è infatti possibile a qualunque utente accedere ai sistemi informativi territoriali di alcuni Comuni e c'è solo da rammaricarsi che ciò avvenga per così pochi. (si veda la Tabella 1)

A questo proposito la Tabella sintetizza il risultato di un'indagine condotta su tutti i capoluoghi di Provincia italiani attraverso Internet per ottenere una mappa dei Comuni dotati di visibilità su *Web* (atteggiamento, questo, già significativo di una certa attenzione al problema ed al rapporto con il cittadino) e, tra questi, dei Comuni che attraverso il*Web* offrono accesso ai propri SIT.

Tabella 1 - Ricognizione attraverso Internet sui comuni capoluoghi di provincia che offrono accessibilità a cartografia comunale

comune	SIT comunale o repertorio cartografico	cartografia interattiva	contenuto informativo	Rete Civica	scarico cartografia
COMUNI CON CARTOGRAFIA PROPRIA ON-LINE (SIT on-line)					
Torino	SIT	si	tecnico, statistico		*
Milano	Urban Exlporer	si	tecnico, statistico e informazioni servizi p.u.	si	
Mantova	SIT	no	tecnico		
Modena	SIT	si	tecnico, statistico e informazioni servizi p.u.		
Bologna	SIT	si	tecnico, statistico e informazioni servizi p.u.		*
Cesena	SIT	si	tecnico, statistico e informazioni servizi p.u.		
Pisa	SIT	no	tecnico	si	
Pesaro	in costruzione			si	
Lecco	in costruzione				
COMUNI CON RINVIO A CARTOGRAFIA NON PROPRIA					
Asti	VIR				
Bergamo	VIR				
Novara	IFD				
Belluno	IFD				
Treviso	IFD				
Pordenone	IFD				
Udine	IFD				
Gorizia	IFD				
Trieste	Rete Civica Trieste	no	Mappa non interattiva, con collegamenti ipertestuali	si	
Imperia	IFD				
Savona	IFD				
La Spezia	IFD				
Lucca	IFD			si	
Pistoia	IFD			si	
Prato	IFD			si	
Siena	IFD				**
Perugia	IFD				
Roma	abcroma	no	Mappa non interattiva, con collegamenti ipertestuali		
Salerno	IFD				
Crotone	IFD				
Catanzaro	IFD				
Vibo Valentia	IFD				
Reggio di Calabria	IFD				
Trapani	IFD				
Sassari	IFD				
Nuoro	IFD				
Oristano	IFD				
Cagliari	Digital Cagliari	no	Mappa non interattiva, con collegamenti ipertestuali		

SIT: Sistema Informativo Territoriale di gestione comunale
IFD: InformaDove http://www.informadove.it
VIR: Virgilio http://www.corenet.it
* scarico cartografia gratuito
** scarico cartografia a pagamento

Innanzitutto è evidente che i 7 Capoluoghi di provincia (9 se includiamo quelli che sono in procinto di consentirlo) che offrono una base cartografica propria sono veramente pochi[8]. Dunque la doppia condizione della presenza di un SIT e della volontà di renderlo largamente accessibile limita a poche unità i Comuni da considerare, a conferma delle preoccupazioni già espresse sul sottoutilizzo del Sistema Informativo.

La consultazione dei siti rivela anche un'eterogeneità terminologica e concettuale molto forte sia nei contenuti che nelle impostazioni metodologiche che richiama l'esigenza della standardizzazione dei formati per agevolare l'interscambio.

La seconda parte della Tabella 1 denota però un nuovo ed interessante atteggiamento dei Comuni: 28 Capoluoghi di Provincia sparsi in 13 Regioni sono dotati di un sito attraverso cui si accede ad una base cartografica comunale *on line* non propria, che fornisce informazioni georeferenziate ("*Informa Dove" o Virgilio* o in alcuni casi un semplice servizio affidato ad Agenzie esterne) grazie ad accordi con altri soggetti (tipici quelli con gestori di archivi come SEAT Pagine Gialle) a dimostrazione dell'importanza assegnata alla disponibilità di una serie di informazioni rivolte al cittadino ma anche dell'incapacità di offrire un proprio SIT. In questi casi si può considerare quella attuale la prima e indispensabile fase di un percorso di modernizzazione della macchina Comunale che richiederà molti e complessi passaggi per conseguire l'effettiva innovazione del rapporto conoscenza - pianificazione - gestione in questi luoghi.

Ciò dimostra un'evoluzione positiva di questi Comuni nella direzione di una risposta alla vastissima ed articolatissima domanda di informazioni georeferenziate per molti usi anche banali, ma la mancanza di risposte a quella domanda di conoscenza "specialistica" a supporto del governo del territorio di cui si è trattato nei primi paragrafi di questo contributo.

Nonostante la grande quantità di dati contenuta in quei veri e propri giacimenti informativi che sono molte Pubbliche Amministrazioni, possiamo sostenere quindi che la circolazione di queste informazioni è ancora scarsa.

Le cause della *ridotta circolazione* delle informazioni sono molte e ci riconducono a problemi già sollevati:

- inadeguata conoscenza dell'esistenza e delle caratteristiche del dato da parte dell'utente (specialista e non specialista) che non è in grado di definire le sue esigenze e di cercare risposte adeguate;
- resistenza dei gestori all'uso ampio dei dati, considerato sovente più un bene privato che un bene pubblico, gravato da procedure burocratiche di accesso complicate e penalizzato da politiche dei prezzi confuse e miopi;
- incoerenza dei dati con i fini auspicati da chi li reclama, sovente sulla base di supposizioni più che di accertamenti.

Questi argomenti rimandano ancora alle due questioni critiche costituite dalla informazione sull'informazione (la *metadocumentazione*) e dall'accessibilità, ma vi aggiungono la necessità di un'azione molto più lunga e complicata su fattori culturali ed istituzionali che

[8] Va precisato che la Tabella 1 non può essere considerata in alcun modo esaustiva, in quanto a questi Comuni andrebbero aggiunti in quelli che hanno adottato iniziative in questa direzione pur senza essere Capoluoghi di Provincia: tra di essi ce ne sono alcuni in cui sono attive interessanti ed innovative esperienze come avviene nel ben noto caso di Carpi.: Per contro vi sono Comuni dotati di importanti SIT che non compaiono in questa lista perché non hanno un proprio sito su Internet (è il caso di Padova e di Venezia, ad esempio).

coinvolga gli utenti (funzionari, professionisti, operatori, tecnici, cittadini) ma soprattutto Enti ed Istituzioni perché si individuino le forme (accordi, convenzioni), i contenuti (archivi, procedure), le modalità (continue, periodiche, con diversi tipi di interscambio) affinché si creino gli auspicati flussi informativi fra i diversi soggetti che solo un'atmosfera di fiducia nella qualità del lavoro altrui e di collaborazione fra partners virtuosi può far diventare efficaci.

Valorizzare il patrimonio informativo dei SIT passa dunque attraverso la possibilità di:

- conoscere a fondo la consistenza del patrimonio informativo, e cioè le informazioni gestite dai singoli settori, il loro contenuto, la loro struttura, i requisiti per la loro acquisizione ed eventuali elaborazioni, ecc. (*metainformazione*)
- definire in maniera chiara le modalità secondo cui queste informazioni possono essere tra loro integrate, valutandone tutti i possibili utilizzi (*istituzionalizzare* le procedure di interscambio).

Solo in questo modo è possibile non solo evitare lo spreco di informazioni e risorse ma, viceversa, aggiungere valore ai dati trasformandoli in informazioni e moltiplicandone le finalità.

3. Difficoltà e condizioni per una gestione efficace del patrimonio informativo comunale a fine di pianificazione.

Molto interesse riveste sotto questo profilo quel particolare tipo di "giacimento informativo" costituito dai tanti archivi di origine amministrativa e tra questi quelli di scala comunale: il riuso di questi dati in molte direzioni non previste dalla loro originale finalità può aggiungere loro del valore significativo soprattutto a fini di pianificazione del territorio. (si veda il Box 1)

Nonostante il dibattito disciplinare da tempo abbia condiviso e sottolineato tutto ciò ed in particolare abbia riconosciuto il carattere strategico di una struttura informativa costruita sulla base delle informazioni raccolte a scala locale, l'utilizzo dei dati amministrativi comunali a questo scopo è ancora alquanto scarso, quantitativamente ma soprattutto qualitativamente: le forme di utilizzo più ricorrenti, infatti, si limitano alle statistiche elementari sulle informazioni più consuete e più generali (tipicamente i dati demografici). Non è raro riscontrare che le stesse Amministrazioni comunali ricorrono, ad esempio in occasione dell'elaborazione di un nuovo Piano Regolatore, ad indagini sul campo (raccolta diretta) per disporre delle medesime informazioni già presenti negli archivi comunali.

I principali fattori da cui dipende l'uso modesto degli archivi amministrativi comunali si possono così sintetizzare:

–la loro organizzazione incoerente con le necessità di chi è incaricato del piano

–la gestione esclusiva che ne fanno i servizi diversi da quello cui compete la pianificazione

–l'incompleta o l'assente informatizzazione e dunque la difficoltà di integrare gli archivi ma anche la lunghezza dei tempi che intercorrono tra raccolta dei dati e successive elaborazioni.

BOX 1

I dati amministrativi comunali[9]

E' ben noto il contributo delle amministrazioni comunali alla statistica "ufficiale"[10]: oltre al supporto in occasione dei censimenti decennali (popolazione, abitazioni, industria, agricoltura, ecc.) gli archivi comunali costituiscono una fonte primaria per le analisi e le verifiche che costituiscono una parte rilevante del supporto conoscitivo nella pianificazione e gestione del territorio, da quelle riguardanti la popolazione a quelle riguardanti l'attività edilizia, l'utilizzo del suolo e degli immobili, la consistenza e l'uso dei servizi pubblici, ecc.. Queste informazioni sono tutte raccolte ed archiviate, anche se nn sempre trattate nei modi e con i criteri più utili per un uso diverso da quello strettamente amministrativo, dalle amministrazioni comunali che devono espletare funzioni proprie dei loro compiti istituzionali attraverso liste di utenti / contribuenti mediante archivi (anagrafe, tributi, licenze di esercizio, concessioni edilizie, utenti di servizi, proprietari di immobili, ecc.) gestiti proprio per tale motivo in modo continuo e puntuale.

Negli archivi comunali dunque confluiscono informazioni su:

- concessioni edilizie;
- lavori pubblici;
- autorizzazioni e licenze;
- occupazione del suolo;
- piano regolatore;
- tributi;
- patrimonio immobiliare pubblico;
- acqua;
- energia elettrica;
- gas;
- telecomunicazioni;
- igiene urbana;
- sanità;
- pubblica sicurezza.

Tutte queste informazioni possiedono un riferimento territoriale ben determinato, attraverso un indirizzo toponomastico o un codice di particella catastale e ciò costituisce per chi si occupa di studiare e governare le trasformazioni del territorio una caratteristica qualitativa di grande rilievo, che molte fonti informative pur fondamentali e ricche non possiedono: lageoreferenziabilità puntuale.

Nel caso di aree urbane dotate di un SIT l'elemento di unione dei vari archivi comunali, attraverso la relazione tra le informazioni e l'unità spaziale cartograficamente individuabile cui esse si riferiscono, è dunque costituito dal codice toponomastico o dal codice di particella catastale (con possibilità in alcuni casi di riferirsi ad entrambi). (si veda il Box 3)

La percezione di questi problemi, ha portato ormai da qualche anno alcuni comuni, e non sempre quelli maggiori ma più sovente quelli di dimensionemedio-grande, ad innovare il proprio modo di trattare le informazioni amministrative introducendo tecnologie informatiche nella loro gestione, sia pure con grande ritardo rispetto a ciò che è accaduto in altri Paesi europei, ma nella maggior parte dei casi lo hanno fatto in modo da produrre ancora modesti effetti qualitativi per le

[9] "Si definiscono dati amministrativi le informazioni raccolte e conservate da istituzioni pubbliche ai fini di controllo o di intervento nei confronti di singoli individui o entità di altro tipo (es. persone giuridiche). Ne sono un esempio quelli raccolti su persone ed imprese ai fini fiscali, pensionistici, anagrafici o giuridici. Il fine, amministrativo, della raccolta di tali dati ne contraddistingue la natura.
Di questi dati è possibile ovviamente fare uso anche a fini statistici producendo indicatori sintetici (frequenze, medie, indici o altro) sugli aspetti relativi alle dimensioni, alle relazioni e alle modificazioni del collettivo di riferimento delle unità analizzate". (ISTAT,http://www.istat.it/metadati/lineeg/22.htm).

[10] Reho, 1992

loro possibilità di governo del territorio. Sono state infatti per lo più informatizzate le procedure gestionali ma si può intuire che solo in pochi casi (come si deduce ancora dalla consultazione della Tabella 1) se ne possono trarre benefici per le analisi di supporto alla pianificazione. I Comuni hanno investito in hardware e software più che nelle competenze necessarie per farne un uso adeguato, ed hanno affrontato scarsamente la riorganizzazione dei flussi informativi interni ed il trattamento automatico dei dati grafici e cartografici.

Altra condizione necessaria per promuovere l'interscambio di informazioni all'interno dell'amministrazione comunale, e per utilizzare al meglio tali informazioni per la gestione del territorio, è la conoscenza precisa e diffusa della disponibilità di informazioni e dunque un loro censimento che consenta di far conoscere ai diversi settori l'esistenza e la disponibilità di queste informazioni, spesso "nascoste" o di difficile reperibilità. al fine di individuare possibili punti di integrazione/interazione ed eventuali disallineamenti tra i database.(si veda a questo proposito quanto esposto più ampiamente al par. 8).

Certamente oggi non sono molti i Comuni che hanno provveduto a censire il proprio patrimonio informativo ed a produrre una *metadocumentazione*[11] aggiornata dei vari archivi in questa direzione.

Al di là delle difficoltà tecniche, del costo e del tempo necessari a compiere quelle che di fatto si configurano come vere e proprie auto-valutazioni sulla gestione della macchina amministrativa, sovente rifiutate perché imbarazzanti non solo per chi ne ha la responsabilità tecnica ma soprattutto per chi ne ha la responsabilità politica, il risultato di una tale operazione comporterebbe riflessi sulla conduzione dell'ente non facili da affrontare né redditizi nel breve periodo e pertanto poco proponibili ad un governo locale il cui arco temporale non può andare oltre gli otto anni di un doppio mandato, al massimo.

L'utilizzabilità (e l'interoperabilità) delle informazioni desunte dagli archivi locali è infatti complicata non solo dai problemi più direttamente connessi con le modalità di raccolta e trattamento dei dati ma, in modo ancor più pesante, da difficoltà di natura organizzativa, culturale e politica.

È possibile sintetizzare tali motivi nei seguenti punti[12]:

- mancanza di una visione *strategica*, nella gestione del patrimonio informativo e debolezza o assenza di una reale correlazione tra le necessità specifiche e quotidiane degli utenti interni ed esterni all'amministrazione;

- modesta o nulla *partecipazione attiva* dei vari soggetti coinvolti nell'interscambio informativo, a tutti i livelli, ed insufficiente impegno concreto nell'adottare forme adeguate di tecnologia informativa per contribuire alla circolazione delle informazioni

- insufficienti condizioni per garantire un buon grado di *stabilità* organizzativa, logistica, politica.

[11] Un interessante esempio di analisi della consistenza del patrimonio informativo comunale è quello bolognese: nel 1995/96, durante le prime fasi di realizzazione del sistema informativo territoriale, l'amministrazione comunale realizzò un'indagine a tappeto per individuare le basi dati gestite dai vari comparti. Il risultato ottenuto fu una lista completa di ben 103 banche dati, contenenti informazioni geografiche ed alfanumeriche provenienti da vari rami dell'amministrazione.

[12] Campbell e Masser, 1995.

Ovviamente, l'analisi e la revisione dell'organizzazione interna dell'amministrazione comunale è la condizione principale per un reale avanzamento nella direzione di una più efficace gestione del patrimonio informativo disponibile: solo attraverso una chiara definizione di ruoli e responsabilità all'interno dei singoli settori è possibile determinare con trasparenza i compiti istituzionali di ciascuno, le regole di interscambio delle informazioni, la messa a disposizione di dati di pubblico dominio, la revisione di procedure amministrative di input e mantenimento dei dati.

E considerazioni non diverse si potrebbero fare sullo stesso problema per il sempre più vasto e costoso materiale informativo prodotto ed archiviato da molti altri Enti (Regione, Enti strumentali, Province, e molti altri Enti pubblici), che si ignorano sia reciprocamente, sia all'interno delle strutture stesse, sebbene possiedano informazioni riconosciute come essenziali per svolgere qualsiasi funzione connessa alla promozione e gestione delle trasformazioni territoriali.

3. Integrare sistemi informativi a supporto del principio di sussidiarietà.

Si pone dunque la necessità di individuare alcune azioni necessarie ed utili per consentire la documentazione, l'accessibilità e l'interscambiabilità degli archivi amministrativi comunali ed aggiungere così valore al patrimonio informativo pubblico, rendendolo più efficace ai fini della gestione del territorio.

Queste azioni si possono così individuare:

- promuovere il superamento di una visione circoscritta all'ambito comunale e proporne, almeno in termini interpretativi, una visione estesa all'ambito corretto dei problemi che quasi mai coincide con la perimetrazione del Comune. Ciò significa rifiutare una visione "*localistica"* dei problemi e la resistenza alla cooperazione. Paradossalmente, infatti, la nuova forza del locale incentivata dalla stessa legge elettorale che oggi consente l'elezione diretta del Sindaco, ha accresciuto il ruolo forte di quest'ultimo, la sua capacità rappresentativa, la sua necessità di interpretare e rispondere in termini diretti, senza la mediazione di partiti nazionali, al suo elettorato. Ciò ha fatto crescere a tutti i livelli la forza delle spinte campanilistiche e rende più difficili la cooperazione orizzontale fra Comuni, l'integrazione verticale con gli Enti territoriali di livello superiore e più in generale ogni iniziativa tesa al coordinamento, a qualunque scala. Il rischio della frammentazione del governo in numerose azioni concertate ma non integrate e della rinuncia, o della impossibilità, di qualunque forma di strategia comune a scala vasta cresce e dunque l'auspicio che si inneschino forme di collaborazione per scambiare informazioni non deve apparire superfluo.

- dare attuazione al principio di sussidiarietà, alle nuove procedure di pianificazione di tipo orizzontale, cooperativo e reticolare, anche attraverso la gestione delle informazioni: proprio la possibilità di assumere totale responsabilità sulle materie che un Ente è in grado di assumersi, comporta la capacità di collaborare con gli altri soggetti e con gli altri livelli

- cogliere l'occasione della elaborazione (o della variazione) di un nuovo PRG per creare o agevolare occasioni di cooperazione fra Enti e fra livelli, ma anche fra Dipartimenti di uno stesso Ente, intorno alla produzione e circolazione delle conoscenze fondamentali (vero momento forte della parte tecnica della pianificazione, come si è argomentato nei primi paragrafi di questo contributo) per rendere la cooperazione reciprocamente vantaggiosa ed attenuare i motivi di diffidenza, la paura di perdere potere e riconoscibilità. E questa necessità non verrebbe meno neanche davanti alle modifiche della normativa urbanistica su procedure e contenuti del piano che possono incentivare l'uso di strumenti diversi dai piani (concertazioni, conferenze dei servizi, accordi di programma): questo patrimonio informativo sarebbe comunque necessario, o lo sarebbe ancora di più, in quanto canale di comunicazione operativo e potenzialmente virtuoso nelle due direzioni (verso i livelli superiori e da questi verso i Comuni) per consentire la registrazione delle variazioni, il monitoraggio della loro attuazione, la valutazione degli effetti fisici, ambientali, sociali. In questa direzione si muovono le esperienze di realizzazione di particolari Sistemi Informativi per il trattamento degli Strumenti Urbanistici che numerose Regioni e Province stanno creando (i "*mosaici*" informatizzati dei piani urbanistici generali) che assumono dati alla scala comunale (il PRG) e li inseriscono in una mappa a scala provinciale e regionale. Qui la collaborazione e l'integrazione sono inevitabili e nel Box 2 si riassumono le caratteristiche di uno studio su questo particolare tipo di sistemi informativi, nella sua applicazione piemontese che appare particolarmente significativa per ragionare su tali aspetti. (si veda il Box 2)

- infine individuare opportune modalità per agevolare l'integrazione fra archivi. Tra queste un ruolo particolare viene svolto dall'uso accorto della georeferenziazione: in particolare su quest'ultima condizione si sofferma l'ultima parte della relazione (si veda il paragrafo 8 e il Box 3).

BOX 2

Il "mosaico dei PRG", un Sistema Informativo mirato alla pianificazione. Suggerimenti da una ricerca sul Sistema Informativo Territoriale per il Trattamento degli Strumenti Urbanistici della Regione Piemonte

Negli ultimi anni molte Regione e Provincie hanno iniziato a realizzare particolari sistemi informativi destinati a memorizzare, pur con modalità diverse, i Piani Regolatori Comunali in forma unificata, puntando ad ottenere alla fine un *mosaico dei PRG* [13] orientato a fornire uno strumento di supporto alle decisioni alla scala territoriale.

E' il caso anche della Regione Piemonte ed il Dipartimento Interateneo Territorio del Politecnico di Torino ha colto più occasioni per collaborare con Regione e CSI-Piemonte[14], alla sperimentazione di metodologie di utilizzo del Sistema Informativo Territoriale per il Trattamento degli Strumenti Urbanistici, ovvero di un elemento rilevante e nuovo del SIT regionale del Piemonte[15] specificamente orientato a far confluire i contenuti del SIT regionale verso le esigenze conoscitive della pianificazione. Grazie alla possibilità di rappresentare in modo integrato le intenzioni di trasformazione dei Comuni dell'area, questo S.I. offre notevoli opportunità e stimoli ad un governo non episodico del territorio: il supporto di immagini territoriali di area vasta ad agglomerazioni, aree metropolitane, territori funzionalmente collegati da aspetti economici, storici, geografici e morfologici comuni ma divisi dalla amministrazione a scala comunale di politiche e piani insensibili ai vantaggi della gestione coordinata, può costituire infatti un utile suggerimento ed un sostegno alle volontà di collaborazione che sovente la società civile manifesta al di sopra dei campanilismi politici.

La ricerca[16] ha anche affrontato il tema della possibile integrazione funzionale tra i diversi livelli decisionali attorno a questo strumento tecnico, soffermandosi in particolare sull'opportunità, di norma poco valorizzata, di utilizzare l'iter e le esigenze connesse alla formazione dei piani comunali come momento strategico per incentivare il rapporto collaborativo fra enti e la circolazione delle informazioni. Lo studio infatti aveva confermato lo scollamento esistente fra Enti a scala diversa: i Comuni ignoravano l'esistenza di un *mosaico dei PRG*; viceversa Regione e CSI-Piemonte non si avvalevano della collaborazione dei Comuni per verificare, correggere, integrare la "trascrizione" dei loro piani. Pertanto sono state messe a punto e sperimentate proposte e raccomandazioni per ottenere questo risultato, valorizzando inoltre il ruolo del *mosaico dei PRG* come depositario di un inconsueto tipo di conoscenza (quella sulle intenzioni di trasformazione espresse nei PRG) e suggerendo la possibilità di farne uso per proporre ad aree funzionalmente integrate a scala sovracomunale una visione più ampia e condivisa di sé stesse e del proprio futuro assetto su cui impostare un governo a scala vasta. Oggi, a due anni dalla conclusione della ricerca, Regione Piemonte e Provincia di Torino, seguendo i suggerimenti conclusivi dello studio, stanno promuovendo, pur con diverse modalità e con azioni separate, la circolazione di informazioni quantitative e cartografiche secondo quelle linee, inviandole a Comuni e Comunità Montane sotto forma di sintetici dossier di "*Documentazione informativa*" contenenti i più significativi elementi quantitativi e grafici utili al Comune tra quelli compresi nel SIT regionale (e tra essi un estratto del *mosaico del PRG* relativo all'area), a supporto tanto della attività comunale di preparazione del PRG quanto della fase di successiva istruttoria per l'approvazione.

Inoltre la diffusione della rete Internet ha aperto una nuova possibilità: quella di consentire a tutti (istituzioni, enti, cittadini) l'accesso all'intero SIT regionale e provinciale e pertanto anche al *mosaico dei PRG* per la parte già attiva.

Non si può ancora parlare di un reale flusso informativo fra Enti nel senso auspicato in questo contributo, ma queste azioni manifestano certamente una sensibilità ed un'intenzione correttamente orientate.

---

[13] I Sistemi Informativi mirati a questo scopo si formano, pur con modalità diverse da Regione a Regione, attraverso l'interpretazione delle cartografie e delle normative di piano, mediante veri e propri *dizionari* che aiutano a tradurre in un numero limitato di voci le variabilissime forme che assumono in ogni PRGC le disaggregazioni di destinazioni d'uso, tipologie di intervento, modalità di attuazione del piano, riconducendo la varietà di modi normativi e cartografici dei diversi PRG ad un formato semplificato e comune, rappresentabile con un *mosaico* dei PRG.

[14] Il CSI-Piemonte è l'ente strumentale della Regione Piemonte che con Politecnico e Università di Torino, Provincia e Comune di Torino, e molti altri Enti locali gestisce il Sistema Informativo Territoriale piemontese.

[15] Ad oggi il Sistema Informativo per il Trattamento della Strumentazione Urbanistica della Regione Piemonte comprende già circa il 50% dei 1207 Comuni del Piemonte tra cui tutti quelli della Provincia di Torino.

[16] Si tratta del *Progetto Mirato B.3. Trattamento degli strumenti urbanistici*, nell'ambito del *Progetto Atlante* finanziato dal Ministero della Funzione Pubblica per promuovere la razionalizzazione del trattamento e della diffusione dell'informazione territoriale e ambientale. Il CSI-Piemonte si è avvalso a tal fine della collaborazione del Dipartimento Interateneo Territorio mediante un contratto di consulenza con Agata Spaziante, responsabile del contratto, e Carlo Alberto Barbieri, concluso con un rapporto dal titolo "Il ciclo informativo intorno alla formazione ed approvazione del PRGC. Problemi e proposte" (CSI-Piemonte, Torino, 1998).

3. Dalla integrazione di SIT comunali informazioni per analisi diacroniche e statistiche a scala subcomunale. Una sperimentazione sul Comune di Torino

A conclusione della relazione, vogliamo portare una dimostrazione convincente delle enormi possibilità di valorizzazione dei contenuti dei SIT locali costruiti a partire da fonti amministrative, riferendo i risultati di una ricerca condotta al fine di verificare la possibilità di ottenere dagli archivi gestionali del Comune (in questo caso il Comune di Torino) elementi conoscitivi utili a monitorare l'attuazione dell'ultimo PRG approvato nel 1995 ed attualmente in una fase di vivace realizzazione di opere previste e promosse dal PRG stesso.[17]

Lo spunto alla ricerca ci è stato dato da una frase contenuta nella relazione sullo Stato di Attuazione del PRG (Città di Torino, 1998) in cui, descrivendo l'evoluzione dell'attività edilizia torinese nel primo triennio dopo l'approvazione del Piano, e riferendosi in particolare agli interventi realizzati nelle aree consolidate (ovvero le aree della trasformazione "ordinaria"), molto meno studiate e seguite di quelle delle ZUT (ovvero Zone Urbane di Trasformazione) oggetto delle operazioni straordinarie di grande portata si diceva quanto segue:

"*Relativamente all'attività edilizia afferente il tessuto consolidato cittadino non disponendo di dati analitici, si è quindi fatto ricorso alle informazioni pubblicate sul Notiziario di Statistica che riporta solo i dati ufficiali relativi alle concessioni edilizie rilasciate limitatamente al numero di nuovi appartamenti […].*

In realtà paradossalmente i dati statistici aggregati sull'attività edilizia tratti dal Notiziario di Statistica, per la citata relazione sullo Stato di Attuazione del PRG, erano stati costruiti a partire dagli stessi dati individuali dell'Archivio delle Pratiche Edilizie gestiti dal Comune stesso e forniti da questo all'ISTAT per le Statistiche sull'attività edilizia e le Opere Pubbliche: caso evidente di viziosa circolazione delle informazioni all'interno della stessa Istituzione.

L'assenza di una sistematica codificazione toponomastica ed informatizzazione dei dati tecnici associati all'archivio di gestione delle pratiche edilizie (volumi,slp, vani, ecc.) aveva fatto sì che ne fosse impraticabile l'uso per monitorare l'attuazione del PRG e per stendere il relativo documento.

Le riflessioni nate da questa circostanza hanno indotto a finalizzare in questa direzione una ricerca sul patrimonio informativo comunale, nel tentativo di mettere a punto (o correggere) e sperimentare alcuni strumenti cui affidare il compito di agevolare (ed in alcuni casi creare) quella circolazione di informazioni fra sistemi informativi locali, ed in particolare fra quelli di origine amministrativa, e fra questi ed i sistemi di altro livello territoriale. Si tratta come già argomentato di una condizione necessaria, sebbene non sufficiente, ad assicurare la possibilità di dare supporto alle strutture di governo a scala vasta ma anche a garantire la qualità del contenuto dei sistemi informativi che gli Enti gestiscono nonché benefici almeno paragonabili al gravoso costo che affrontano per produrli e mantenerli attivi.

A partire da una prima analisi dei caratteri e dei contenuti dei principali archivi prodotti dalle azioni amministrative del Comune di Torino:

- Pratiche edilizie

[17] La ricerca è stata condotta presso il Dipartimento Interateneo Territorio da Piergiorgio Cipriano, con una borsa di studio del CSI-Piemonte. Vi hanno collaborato per il Dipartimento Interateneo Territorio Agata Spaziante e Franco Vico, per il CSI-Piemonte Adriano Sozza e Marco Cavagnoli. Ha collaborato il Comune di Torino.

- Anagrafe degli immobili
- Toponomastica
- Archivio Imposta Comunale sugli Immobili (ICI)
- Tassa raccolta Rifiuti Solidi Urbani e Tributo Ambiente (TARSU).

nonché dell'archivio del

- Catasto Edilizio Urbano

si sono studiate le possibilità per alimentare il rapporto fra questi Enti o loro parti producendo informazioni nuove e trattamenti nuovi dallebasi-dati ordinarie per alimentare un vero flusso, biunivoco e duraturo, di informazioni sullo stato e sulle prospettive di cambiamento del territorio, nella convinzione che ciò possa anche contribuire a ridurre la conflittualità e le incomprensioni che oggi caratterizzano, come argomentato nei primi paragrafi di questo contributo, i rapporti fra dipartimenti di una stesso ente, forse ancor più che fra enti diversi, compromettendo ogni possibilità di cooperazione e coordinamento.

Si sono così messi a punto e sperimentati operazioni di integrazione fra questi archivi gestionali per verificare la produzione di nuovi livelli informativi (quello che abbiamo considerato il "valore aggiunto informativo").

La ricerca ha posto al centro della sua attenzione l'archivio delle pratiche edilizie ed ha innanzitutto confermato l'importanza che assumono i due aspetti già indicati come le criticità prioritarie su questo problema:

- l'omogeneo ed istituzionale riconoscimento del *codice toponomastico* come indispensabile collegamento delle pratiche amministrative con il trattamento delle informazioni ad altri fini (statistici e di pianificazione in particolare) (si veda il Box 3)
- la *metadocumentazione* degli archivi

Si è voluto e potuto dimostrare che i dati di natura amministrativa rintracciabili attualmente nell'archivio delle pratiche, e già privi di una documentazione adeguata, qualora se ne completasse la georeferenziazione attraverso l'indirizzo oggi incompleta, potrebbero essere utilizzati per interessanti analisi ripetibili nel tempo sull'evoluzione del patrimonio edilizio e dell'attività edilizia, oggi improponibili.[18]

Naturalmente il carattere gestionale dei citati archivi comporta non pochi problemi nel selezionare le informazioni utili e coerenti con le finalità di un'analisi sull'evoluzione edilizia della città. Ad esempio considerare la totalità dei provvedimenti di concessione rilasciati negli anni, significa avere a che fare con una miscellanea di opere più o meno rilevanti, che vanno da chiusure di intercapedini, a costruzione di nuovi edifici, ampliamenti, cambi di destinazione d'uso, modifiche esterne, ecc. poiché le informazioni tecniche non vengono trascritte (per il momento) nell'archivio informatizzato di gestione dell'iter e rimangono all'interno della documentazione cartacea presentata dal progettista. Pichè ai fini del monitoraggio dell'attuazione del PRG è soprattutto importante considerare la distribuzione nello spazio e nel tempo di quelle

[18] L'archivio delle pratiche edilizie non è completamentegeoreferenziato ma si è potuto facilmente localizzare il 90% degli interventi edilizi di cui esista pratica

BOX 3

Il codice toponomastico come elemento di georeferenziazione dei dati amministrativi comunali. Due stili a confronto.

L'opportunità di scambiare informazioni e connettere archivi richiede che vengano individuati opportuni *link* fra archivi per agevolarne l'integrazione.

A livello comunale la rappresentazione spaziale delle informazioni amministrative utilizza preferibilmente l'indirizzo (e quindi il codicetoponomastico) come riferimento spaziale del dato. Pertanto la gestione dell'archivio Toponomastica all'interno dell'amministrazione comunale risulta essere per i settori amministrativi e tecnici un fattore di primaria importanza, che gli appositi uffici in genere gestiscono senza procedure molto "rigide", nonostante ne sia riconosciuta l'ufficialità a livello amministrativo. Pertanto il modo in cui la relazione pratica - indirizzo viene costruita ed utilizzata è molto variabile da Comune a Comune.

Qui si citano due esempi di diversi "stili" nell'uso di questo fondamentale riferimento geografico dei dati comunali in due casi (Torino, Modena) significativi della riflessione che si vuole proporre: si tratta infatti di Comuni dotati di importanti SIT comunali e di rilevante esperienza e tradizione nell'affrontare il problema della informatizzazione della pubblica amministrazione e dunque di casi già particolarmente fortunati.

Nel caso della città di Torino l'archivio della Toponomastica è gestito del Settore Statistica, in sovrapposizione con i Settori Anagrafe e Tributi ed in collaborazione con la Divisione Urbanistica e Edilizia. L'utilizzo di questo codice come elemento di connessione fra archivi e procedure comunali, sia amministrative che tecniche, pur essendo largamente utilizzato, non è ancora istituzionalmente riconosciuto, per esempio, all'interno dell'iter delle pratiche edilizie dell'archivio "pratiche edilizie", qui privilegiato perché utilizzato nella sperimentazione sul trattamento di dati relativi all'attività edilizia del Comune (si veda il par. 7). Non esiste pertanto alcuna procedura formalizzata di comunicazione, aggiornamento e controllo puntuale nei casi di nuove costruzioni e/o demolizioni, e dunque si perde la possibilità di un controllo efficace dell'urbanizzazione, come in altri Comuni (es. quello illustrato di seguito).

Nel caso della città di Modena l'ufficio Toponomastica, situato presso il Settore di Pianificazione Territoriale insieme all'ufficio di cartografia-topografia, identifica la codifica toponomastica con l'anagrafe immobiliare, in quanto, già a partire dalla revisione avvenuta nel 1986, le singole unità immobiliari (residenziali e non) costituiscono l'unità spaziale di riferimento degli archivi comunali. La numerazione civica domina di fatto la pratica quotidiana dei vari uffici comunali vincolando a questa codifica l'avvio di qualunque pratica amministrativa. Ciò consente l'immediata georeferenziazione di dati relativi a soggetti (fisici o giuridici) e oggetti (immobili), ed il loro agevole inserimento nel SIT del Comune: Inoltre, la forte sinergia con la locale azienda municipalizzata di distribuzione di energia (META) permette di individuare e codificare in via preventiva ogni nuova unità insediativa, indipendentemente dal fatto che la localizzazione sia passata attraverso una forma di autorizzazione comunale. Sono evidenti le ricadute operative per l'amministrazione della macchina comunale, per il controllo dell'urbanizzazione e dell'abusivismo edilizio ma anche per la produzione di conoscenze utili alla pianificazione del territorio ed al monitoraggio dell'attuazione del PRG (ad esempio attraverso l'osservazione" dell'attività edilizia che fa seguito all'entrata in vigore di un nuovo PRG).

Può essere interessante a questo proposito esaminare le tracce che, in questo Comune, la pratica di una nuova costruzione lascia nella gestione comunale, generando un nuovo punto codificato. Una pratica edilizia, quale il rilascio di una concessione ad un intervento edilizio, innesca dunque una mutazione del tessuto urbano ma, in questo modo, attiva anche l'aggiornamento degli archivi quantitativi e cartografici. Già alla richiesta di allacciamento per cantiere effettuata dal costruttore del nuovo fabbricato viene attribuito un nuovo numero civico in forma provvisoria e sintetica (l'unità spaziale è l'intero fabbricato in via di costruzione).A sua volta questo sarà composto da singole unità immobiliari già individuate, in fase di richiesta di concessione, nell'apposito elaborato di sintesi presentato con la documentazione di rito al Settore Gestione e controlli - Servizio Trasformazioni Edilizie. In tale documento (schema grafico di sintesi) è riportata la suddivisione prevista dal progetto nelle singole unità ecografiche. Successivamente, con la richiesta di allacciamento alla rete elettrica, per le singole unità immobiliari viene trasmessa comunicazione all'ufficio toponomastica, il quale provvederà all'attribuzione dei numeri interni relativi alle abitazioni, ai locali destinati ad attività commerciali, artigianali, uffici, box auto, depositi, ecc.

Il documento grafico sintetico, unito alla tabella dati delle unità funzionali oggetto dell'intervento, fa da guida per la corretta attribuzione della numerazione civica interna degli edifici. In questa fase, inoltre, l'ufficio Toponomastica ha il compito di aggiornare anchecartograficamente la numerazione civica, introducendo il nuovo elemento (provvisorio)[i] a cui l'ufficio cartografico del Settore Pianificazione farà corrispondere l'introduzione in mappa della sagoma provvisoria dell'edificio in costruzione.

La condizione di "provvisorietà" del nuovo edificio mappato durerà fino al momento dell'avvenuto accatastamento, richiesto per il rilascio del certificato di abitabilità/agibilità. A quel punto il codicetoponomastico diventerà definitivo.

opere aventi un peso maggiore (in termini di nuova cubatura, oneri di urbanizzazione, carichi urbanistici) si è proceduto a selezionare soltanto i casi di aumento di volumetria, legati a nuove costruzioni, ampliamenti o sostituzioni attraverso il modello ISTAT allegato alla pratica.[19]

Restringendo in questo modo il set ai soli casi di concessioni rilasciate annualmente, si può ottenere la sintesi quantitativa contenuta nelle Tabelle 2 e 3.

ANNO	APPARTAMENTI CONCESSIONATI
1994	649
1995	1042
1996	1449
1997	1187
1998	1173 (dato riferito al mese di ottobre)[20]

Tabella 2– Concessioni rilasciate per anno

	1994	1995	1996	1997	1998	1999
nuovi edifici	60	41	60	75	66	82
ampliamenti	40	52	53	65	37	38
box/autorimesse	66	47	47	38	50	66
bassi fabbricati	13	32	11	9	11	10
ristrutturazione	35	27	63	65	65	51
conservazione	54	32	84	23	35	39
modifiche	133	30	75	86	73	58
cambio d'uso	73	24	36	44	58	90

Tabella 3 – Concessioni rilasciate per tipi di intervento

ma la georeferenziazione consente anche una rappresentazione grafica delle concessioni per edilizia residenziale rilasciate di anno in anno, come quella riportata nella Figura 1.

La sintesi a scala urbana di questi dati, ovviamente ottenibile anche a scala subcomunale fino al massimo livello dettaglio, consente interessanti considerazioni sulla quantità e distribuzione dell'attività edilizia in rapporto alle attese del novo PRG (non va trascurato il fatto che il PRG è entrato in vigore nel 1995), alla stasi seguita alla sua approvazione ed alla ripresa tutt'altro che omogeneamente distribuita nel tempo e fra le aree della città nei primi cinque anni della sua azione di governo del territorio. Diventano dunque possibili analisi diacroniche e per parti di città sui processi locali, ripetibili con grande facilità.

Non meno interessanti osservazioni, verifiche, simulazioni si possono fare e si potranno illustrare nel corso del Workshop, intersecando questi dati con la cartografia di PRG e dunque

[19] Va precisato che per questa via non vengono inclusi i dati che fanno riferimento a ristrutturazioni, conservazioni e cambi di destinazione d'uso e che rappresentano, in particolare per il tessuto consolidato, una parte considerevole.

[20] I valori riportati comprendono la globalità delle concessioni rilasciate, riferite a tutte le trasformazioni (sia quelle "ordinarie" sia quelle nelle Zone Urbane di Trasformazione - ZUT)

con quel *Sistema Informativo per il Trattamento degli Strumenti Urbanistici* o *mosaico dei PRG* illustrato nel Box 2, con il Catasto Edilizio Urbano, con l'Anagrafe comunale.

E considerazioni particolarmente interessanti si sono fatte anche tenendo conto che il codice toponomastico costituisce chiave relazionale con le sezioni di censimento e quindi con quel grande patrimonio di dati che sono i passati censimenti: si apre la possibilità di integrare e comparare dati locali con dati statistici nazionali come quelli che vengono dalle fonti informative statiche costituite dai censimenti. Ciò apre prospettive di grande rilevanza anche sull'uso futuro di questo genere di dati con quelli che il prossimo anno l'ISTAT aggiornerà.

Dunque si è potuta dimostrare la fattibilità e l'utilità di un'integrazione che consente di aggregare al valore proprio dei dati di questo archivio destinato in origine a gestire l'attività edilizia e l'iter delle autorizzazioni relative, un valore informativo addizionale che stimola l'opportunità di trarre significati analitici e statistici utili in diverse direzioni.

3. Conclusioni

Le argomentazioni ed i risultati del lavoro applicativo condotto sul caso torinese ribadiscono l'importanza di un'azione per diffondere la sensibilità di tecnici, professionisti, amministratori nei confronti del problema proposto nei primi paragrafi di questa relazione: di fronte ad una crescente e sempre più articolata domanda di conoscenza, ed in particolare di conoscenza a supporto del governo del territorio nelle sue mutevoli e sempre più complesse forme, le pubbliche amministrazioni non possono più consentirsi disottoutilizzare, se non di sprecare, il rilevante patrimonio informativo che contribuiscono ad accumulare con la propria attività gestionale.

Si impone la necessità di scoprirne le potenzialità non sfruttate, di accrescerne la qualità, di renderne ampiamente utilizzati (e dal più vasto numero possibile di utenti) i contenuti.

Le caratteristiche sofisticate che si richiedono in particolare alle informazioni idonee ad affrontare i nuovi connotati che piani e politiche stanno assumendo (processualità, trasparenza, equità distributiva, continuità, flessibilità, interscambiabilità) impongono alle amministrazioni comunali non solo l'urgenza di introdurre nella organizzazione del lavoro le nuove tecnologie informatiche strutturando i flussi informativi in Sistemi Informativi Territoriali, ma l'importanza di procedere oltre, connettendo procedure, archivi, sistemi informativi frammentati fra Enti e fra parti di Enti per integrarli.

Solo in questo modo è possibile l'uso esteso e la valorizzazione di quei giacimenti informativi che oggi restituiscono modesti benefici in cambio degli alti costi di formazione e manutenzione che richiedono.

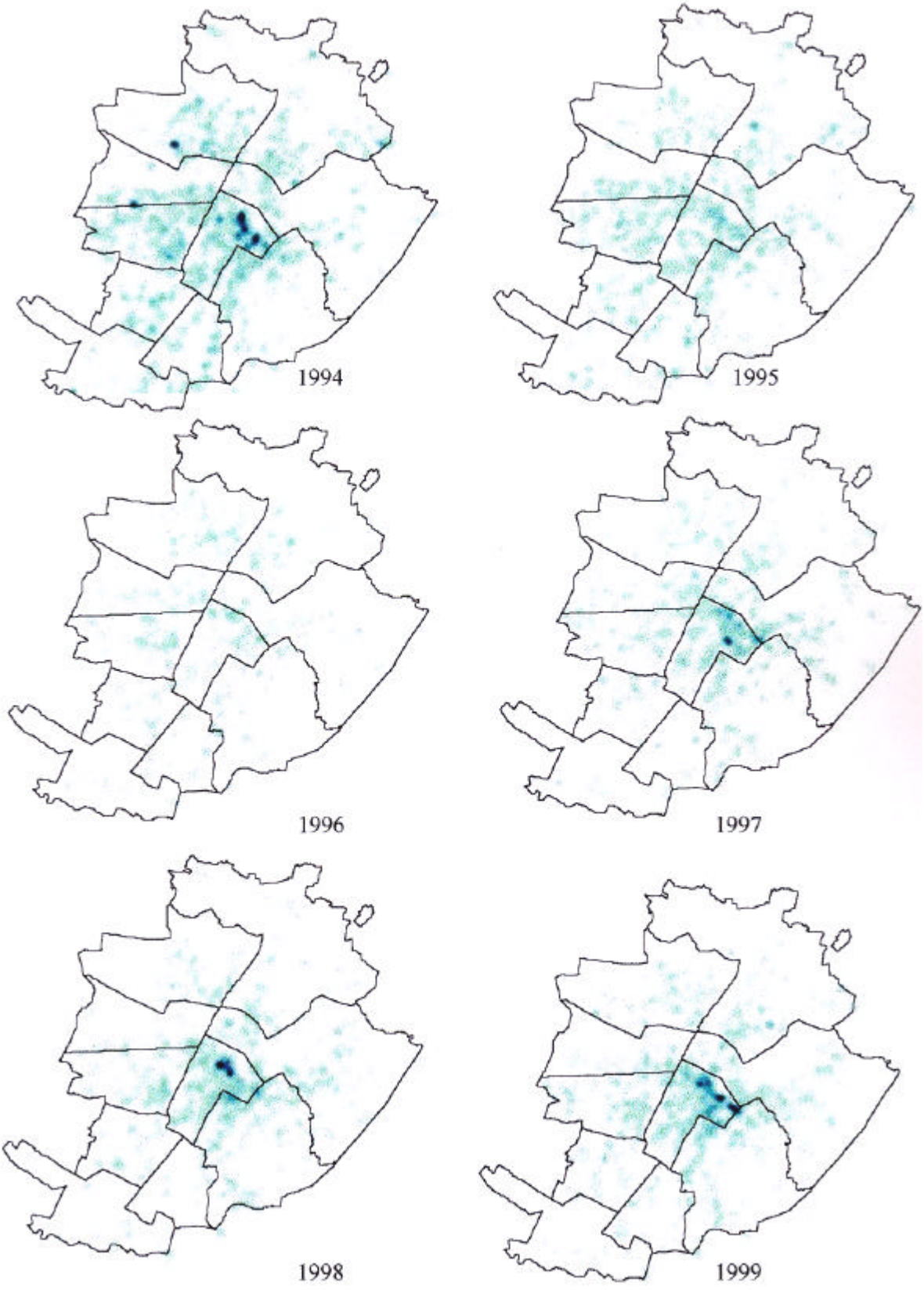
A questo fine è necessario puntare soprattutto su due azioni:

- l'uso multiplo dei dati, attraverso una opportuna strategia digestione del gestione del patrimonio informativo a livello comunale;
- l'incremento della conoscenza e dell'accessibilità alle informazioni, per poterne garantire una maggiore diffusione e condivisione.

Le ricerche condotte dal Dipartimento Interateneo Territorio del Politecnico di Torino con il CSI-Piemonte per approfondire questi problemi e sperimentare delle soluzioni, hanno in molti modi documentato questi aspetti, a conferma delle ipotesi formulate in partenza, e sperimentato la

fattibilità di azioni non particolarmente complesse ed onerose attraverso cui il valore informativo aggiunto sembra presentarsi come particolarmente interessante.

Figura 1 - Comune di Torino. Concessioni edilizie rilasciate nel periodo 1994-1999



Hanno infatti documentato l'ancora modesta diffusione dei SIT presso i maggiori Comuni (i capoluoghi di Provincia) e per contro il diffondersi di una certa attenzione di questi Comuni alla visibilità consentita dalla apertura di siti attraverso cui offrire l'accesso ad informazionigeo-referenziate per gli usi più vari. Fra questi si è anche segnalato ila particolare rilevanza che può assumere l'integrazione con un particolare SIT orientato a sostenere la pianificazione a scala vasta com'è il Sistema Informativo per il Trattamento degli Strumenti Urbanistici ovvero il *mosaico dei PRG,* particolarmente interessante perché consente ai soggetti della pianificazione di disporre di un quadro di riferimento complessivo delle intenzioni di trasformazione espresse dai Comuni nei PRG e di promuovere la circolazione di informazioni fra Comuni, Provincie, Regione intorno all'iter di formazione del PRG.

La parte esemplificativa ha anche sondato e confermato la fattibilità di un uso a fini conoscitivi degli archivi gestionali del Comune di Torino pemonitorare l'attuazione del PRG, per produrre analisi diacroniche sull'attività edilizia, per alimentare ed integrare statistiche sulla produzione edilizia a partire dagli archivi delle pratiche edilizie.

I risultati delle applicazioni confermano che alcune condizioni si presentano come strategiche per ottenere questi progressi:

- comune a tutti è il problema della carente conoscenza (all'interno dell'amministrazione) del proprio patrimonio informativo e cioè delle informazioni gestite dai singoli settori (contenuto, struttura, requisiti per l'acquisizione): in una parola *metainformazione;* è necessario che si proceda ad analisi dettagliate degli archivi dispersi all'interno delle amministrazioni comunali, in maniera normalizzata e programmata;
- un secondo nodo di fondamentale importanza per la *georeferenziabilità* delle informazioni di questo tipo è la sistematica utilizzazione del *codice toponomastico*, ovvero dell'indirizzo anagrafico; anche a questo proposito esistono profonde discrepanze tra i vari comuni (per storia, cultura, morfologia del territorio) che occorrerebbe eliminare o ridurre attraverso linee-guida di carattere nazionale
- una terza condizione è costituita dalla standardizzazione dei formati di scambio dei dati sia all'interno delle Istituzioni sia verso l'esterno (comuni contermini, Provincia, Regione) istituzionalizzando le procedure d interscambio con l'obiettivo di usare al meglio e di riusare, anche a fini statistici, la miriade di dati raccolti quotidianamente;
- una quarta condizione è costituita dalla normalizzazione, con indicazioni specifiche di carattere nazionale, delle modalità di accesso, diffusione e visibilità delle informazioni contenute nei SIT comunali (accesso gratuito, a pagamento, con convenzioni, ecc.)

Queste condizioni, necessarie ma non sufficienti, devono costituire una priorità nell'azione e negli interventi di amministrazioni centrali e periferiche per permettere di coordinare trattamenti, scambiare informazioni, standardizzare formati, nell'ottica della realizzazione di reti di SIT:

E' questo il percorso da privilegiare per valorizzare il patrimonio informativo giacente presso le pubbliche amministrazioni e ricevere un supporto efficace per la produzione di conoscenza destinata al governo del territorio.

Bibliografia

Aigotti D., Cavagnoli M., Garretti L. (1995), "Il sistema informativo per il trattamento della strumentazione urbanistica della Regione Piemonte",*Documenti del Territorio*, n. 30/31.

Artuffo G., (1999), Rassegna Stampa, in MondoGIS - Rassegna Stampa, http://www.domino-research.it/rass_stampa.htm

Barbanente A. (1996), "La pianificazione di area vasta. Esperienze regionali e provinciali" (prima parte), dossier 3/96, supplemento al n. 149 di *Urbanistica Informazioni*.

Caiazzo L., http://ordingsa.general.it/rivista/rivista.htm#num21

Campbell H.J., Masser I. (1995), *GIS and organisations: how effective are GIS in practice?*, Taylor & Francis, London

Cavagnoli M., Garretti L. (1996), "Sistemi informativi e pianificazione urbanistica: il sistema informativo per il trattamento della strumentazione urbanistica della Regione Piemonte", Relazione presentata alla *XVII Conferenza Italiana di Scienze Regionali*, Sondrio, volume 2, 529-541.

Cavalcoli P. (1998), "I Siti delle province; venti possibili progetti", *Workshop sui Sistemi informativi territoriali*, Bologna 26-27 febbraio 1998, (dattiloscritto).

Città di Torino (1999), *Stato di Attuazione del PRG al 31/12/1998 - Comparto produttivo e artigianale - Studio di Variante al PRG*, Torino

Chiabodo R., Vico F. (1999), *Stato di avanzamento della Città - Rapporto Preliminare di sintesi dell'incontro con i funzionari della Divisione Edilizia e urbanistica - Settori trasformazioni convenzionate e Strumentazione urbanistica*, Torino

Ciancarella L., Durazzi A., Secondini P., Valpreda E. (1993), "Geographical Information System in enviromental analysis and planning: reasoning trough a review of the principal european and italian experiances", *Proceeding of Egis '93 Conference*, Genova 29 marzo - 1 aprile 1993, vol. I, pp. 555-564.

Comune di Modena, http://sit.comune.modena.it

Delpiano A. (1998), "Applicazione operative dei Prg informatizzati all'interno del sistema informativo Territoriale", *Workshop sui Sistemi Informativi territoriali…*, op. cit. (dattiloscritto).

INU (1995), "La nuova legge urbanistica. I principi, le regole", *XXI Congresso nazionale*, Bologna.

ISTAT, http://www.istat.it/metadati/lineeg/22.htm

Laurini R. (1996), "Stato dell'arte dei SIT e prospettive per la pianificazione regionale e urbana", Relazione presentata alla *XVII Conferenza Italiana di Scienze Regionali*, Sondrio, volume 3, 571-582.

Maguire D.J., Dargermond J. (1991), "The funtionality of Gis", in Maguire D.J., Googchild M.F., Rhind D.W., *Geographical Information System*, Essex, pp. 29-50.

Reho M., (1992), "Fonti informative per l'analisi territoriale", *Collana R*, num. 12, DAEST, Venezia

Scarduelli G. (1995), "Il sistema informativo territoriale come supporto alla pianificazione d'area vasta", in *Urbanistica Informazioni*, n. 5, 1995, pp. 90-95

Sartorio G., Spaziante A. (eds.) (1995), *Il piano regolatore generale nella legge urbanistica del Piemonte*, Politecnico di Torino, Regione Piemonte.

Seassaro L. (1996), "Le dimensioni organizzative dei sistemi informativi. I SIT nella prospettiva della pianificazione come processo continuo a carattere sociale e interattivo", Relazione presentata alla *XVII Conferenza Italiana di Scienze Regionali*, Sondrio.

Spaziante A. (1996), "L'esperienza di cartografia informatizzata della Regione Piemonte", in Di Biagi P., Ernesti G., Gabellini P., Mazzoleni C., Treu M.C. (ed), *La rete archivi dei piani urbanistici, Triennale di Milano*.

Spaziante A.(1998), "Dal piano al sistema informativo e ritorno", *Urbanistica Informazioni*, n. 159, pp. 19-20.

Spaziante A. (2000), "La cooperazione istituzionale per il governo del territorio",Archi@Media, pp. 61-79.

WORKSHOP

Sistemi informativi territoriali

Le infrastrutture dell'informazione geografica: cooperazione e condivisione dei dati geografici di base

F. Vico

Le infrastrutture dell'informazione geografica: cooperazione e condivisione dei dati geografici di base

Franco Vico

Dipartimento Interateneo Territorio del Politecnico e dell'Università di Torino

Sommario
Le finalità di un ente quale l'ISTAT sono produrre e distribuire l'informazione di base, da utilizzare con diverse finalità in vari campi. Collocando queste finalità nel contesto attuale, caratterizzato dallo sviluppo di Internet, dalla diffusione dei sistemi informativi territoriali ecc., il concetto di infrastruttura di dati geografici può diventare di riferimento sia nella produzione che nella distribuzione delle informazioni di base.

Gli elementi costitutivi di una infrastruttura di dati geografici sono:
- il catalogo unificato dei dati;
- i metadati, la documentazione standardizzata dei dati necessaria per valutare se servono per i propri scopi e per poterli utilizzare;
- i dati di base perché utilizzati da molti o perché di riferimento per altri georeferenziazione di altri dati.

Sullo sfondo c'è Internet, che è lo strumento per accedere al catalogo dei dati e ai dati.

L'approccio "infrastruttura di dati geografici" implica di vedere l'informazione (geografica), appunto, come una infrastruttura, indispensabile come la rete stradale, e di vedere i flussi informativi tra i diversi soggetti non come risultato di specifici accordi ma come cooperazione e condivisione di dati in un contesto più ampio, in cui, possiamo dire, domanda e offerta di informazione geografica si confrontano.

La dizione "infrastruttura di dati geografici" è stata introdotta negli USA intorno al 1994: il punto 2 del *paper* è appunto dedicato ad una interpretazione dell'esperienza americana della National Spatial Data Infrastructure (NSDI), che oggi è una realtà, ancorché *in progress*.
In Europa la dizione infrastruttura di dati geografici appare subito dopo, nel 1995: tra gli obiettivi della European Commission c'è la costituzione della European Geographic Information Infrastructure (EGII). Ma i risultati concreti conseguiti sono ancora relativamente scarsi. I motivi sono diversi, da ricercare nella frammentazione normativa dell'Europa, nella molteplicità e diversità dei soggetti che nei differenti paesi producono dati geografici. Il punti 3 è dedicato a dare un quadro dello stato dell'arte per quanto riguarda la costruzione della EGII. Il punto 4 è dedicato all'idea di una Global Spatial Data Infrastructure, e il punto 5 al progetto italiano di Sistema di Comunicazione dati territoriali (SCT). Nel punto 6 sono tratte alcune conclusioni: mettendo insieme il concetto di infrastruttura di dati geografici e webGIS, la tecnologia per visualizzare e interrogare interattivamente dati geografici attraverso Internet, viene delineata la prospettiva di un un "*societal or internet GIS"* diffuso su Internet, su più livelli, che coinvolge produttori e utilizzatori di dati geografici.

1. Condivisione dei dati geografici: un modo troppo strutturato e uno troppo poco

Un'esperienza che ho fatto più volte in questi ultimi dieci anni, direttamente o indirettamente (seguendo lavori di tesi di studenti) è quella di cercare di usare dati di varia natura, raccolti e organizzati da un ente locale in relazione a propri compiti istituzionali, a fini di ricerca.
Più o meno la storia è stata sempre la stessa. All'inizio c'è una vaga notizia di dati che potrebbero essere utili. A questo punto si comincia con le telefonate e le lettere, per ottenere maggiori informazioni e le necessarie autorizzazioni. Quando, dopo un bel po' di tempo (generalmente), si arriva, se si arriva, al soggetto che detiene il controllo materiale dei dati, è possibile che questi vengano forniti oppure no.
Nel primo caso, è improbabile che i dati vengano accompagnati dalla documentazione necessaria, cioè vengano fornite spiegazioni su come i dati sono stati prodotti e, a volte, neppure su come decodificare con certezza il significato dei campi...
Se i dati ci sono ma non vengono forniti, ciò non è quasi mai dovuto al fatto che sono realmente riservati. Spesso il motivo (non dichiarato ovviamente) è che la qualità dei dati è così bassa, i dati sono così lacunosi e male organizzati dal punto di vista informatico, da essere impresentabili.
La condivisione dei dati (se avviene) è costruita di volta in volta, ogni volta spendendoci non poco tempo. Non c'è quasi mai cooperazione fra i diversi soggetti, nel senso che non c'è quasi mai un ritorno per chi fornisce i dati: questo forse giustifica, almeno in parte, la riluttanza a fornire i dati.

Se questo modo di condivisione dei dati geografici è del tutto destrutturato, troppo strutturato è quello basato sul concetto *data-warehouse* (DWH).
Un DWH[1] integra, sommarizza, generalizza[2] dati provenienti da diverse fonti: l'insieme di queste operazioni può descritto come una *"successione ordinata di funzioni applicate ad un set di dati in input con la conseguente produzione di un output"* (Di Pace, Nicoletti, 2000, p. 742).
I dati gestiti in un DWH sono (SAS, 1995):

- *subject-oriented*, cioè organizzati in relazione ai differenti modi in cui gli utilizzatori li vedono e li usano;
- codificati in modo consistente, mentre dati prodotti e gestiti in ambienti diversi possono generare inconsistenze quanto ne è richiesta l'integrazione;
- organizzati in modo consistente rispetto al tempo, cioè in modo da permettere comparazioni temporali;
- non-volati, una volta che sono inseriti nel DWH non sono più aggiornati o cambiati.

I produttori di sw presentano il DWH come uno specifico sw o come un set di strumenti sw. Certamente un DWH deve includere strumenti di query, di data-analysis, applicazioni OLAP (*on line analytical processing*) ecc., ma le questioni chiave sono la sua progettazione logica e gli aspetti organizzativi. I punti di partenza nella progettazione del DWH sono:

- la comprensione dei fabbisogni di informazione (presenti e futuri) dei suoi utilizzatori (fondamentalmente i *decision-makers*),
- l'analisi degli attuali flussi di dati.

---

[1] Il concetto di *data-warehouse* è stato proposto con enfasi in particolare da SAS (il produttore del ben noto sw statistico) a partire dalla metà dagli anni '90, ma era già usato da IBM fin dal fine degli anni '80. Le sue varianti geografiche sono *geo-data-warehouse* (Thewessen and van de Velde, 1995, p.78) e *spatial- data- warehouse,* dizione introdotta nel 1997, in specifico da ESRI.

[2] Mi piace qui usare il termine generalizzazione, usato in cartografia per indicare il processo che deve essere applicato nel momento in cui si deriva una carta in scala minore da una base in scala maggiore e quindi si deve passare da una rappresentazione in scala (ad esempio della struttura viaria di un centro urbano) ad una simbolica: è giusto dire che è un'arte oltre che una tecnica.

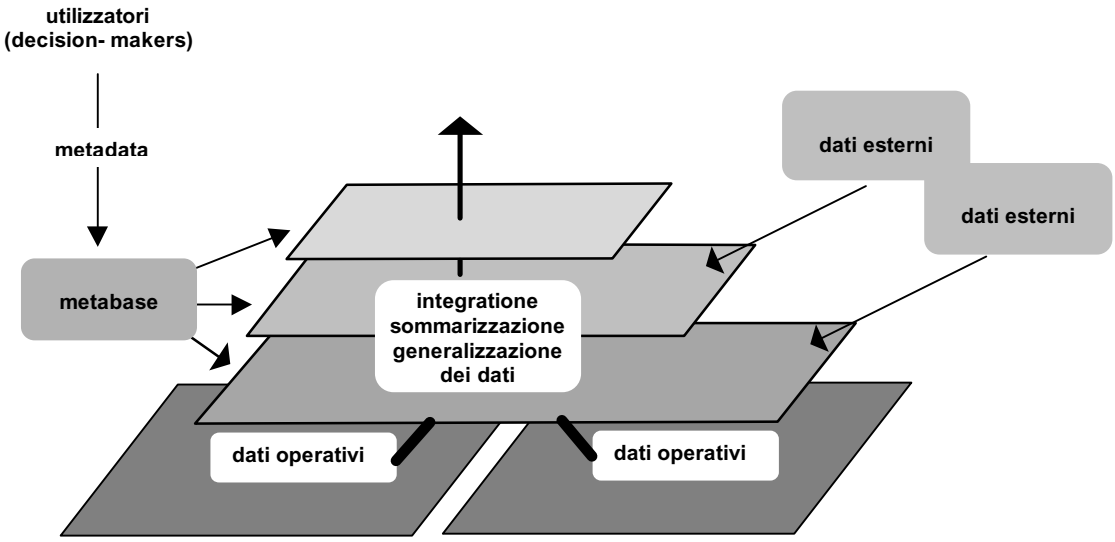


Figura 1. Il concetto di data-warehouse[3]

Se l'obiettivo è produrre e distribuire l'informazione geografica di base, da utilizzare con diverse finalità in vari campi (che è l'obiettivo con cui mi sto confrontando in questo *paper*) questo modo di organizzare la condivisione dei dati e la cooperazione può risultare troppo rigidamente strutturato, anche per l'enfasi che viene data alla precisa individuazione della domanda di informazione di determinati soggetti.

2. La National Spatial Data Infrastructure americana

Il modo più semplice per introdurre il concetto di infrastruttra di dati geografici è quello di discutere l'esperienza americana.

2.1 L'Executive Order del 1994

«Le informazioni di tipo geografico sono cruciali nel promuovere lo sviluppo economico, nel migliorare l'uso consapevole delle risorse naturali e nel proteggere l'ambiente» (Executive Order, 1994). Questa giudizio apre l'Executive Order del Governo Federale degli Stati Uniti che istituisce la National Spatial Data Infrastructure (NSDI) per coordinare la produzione e l'accesso ai dati geografici.

> *«"National Spatial Data Infrastructure" (NSDI) significa la tecnologia, le politiche, gli standard, le risorse umane necessarie per acquisire, trattare, memorizzare, distribuire e migliorare l'utilizzazione dei dati geospaziali.»*
> (Executive Order 12906, 1994)
>
> *«"Dati Geospaziali" significa le informazioni che identificano la localizzazione geografica*

[3] Il concetto di *metadata* viene spiegato nel punto successivo

e le caratteristiche di elementi naturali o costruiti o linee di confine esistenti sulla terra. Queste informazioni possono derivare, fra l'altro, da telerilevamento, carte, rilevamenti sul terreno. I dati statistici possono essere inclusi in questa definizione a discrezione dell'ente che li raccoglie.» [4]
(Executive Order 12906, 1994)

L'Executive Order consolida un progetto avviato nel 1989. La *leadership* nella costruzione della NSDI è affidata al Federal Geographic Data Committee (FGDC), un organismo istituito nel 1990, che comprende 17 agenzie federali, nel quale sono presenti anche stati, governi locali, università e il settore privato.
Gli scopi sono molto concreti (diminuire gli sprechi e evitare la duplicazione delle iniziative) ma nel contempo c'è una visione ampia dell'informazione geografica e delle attività ad essa connesse: il processo che viene attivato è di lunga durata, comporta l'uso di ingenti risorse, e dovrà coinvolgere a un certo punto le 80.000 agenzie pubbliche che negli Stati Uniti trattano in qualche modo informazione geografica (Masser 1998, p 88).
Le componenti chiave di questo progetto sono tre:
- la "National Geospatial Data Clearinghouse";
- i *metadata*;
- i *framework data.*

Due altre importanti componenti sono:
- Internet,
- il quadro istituzionale.

2.2 La Clearinghouse

«"National Geospatial Data Clearinghouse" significa un network distribuito di produttori, gestori, utilizzatori di dati geospaziali connessi elettronicamente.»
(Executive Order, 1994)

«Usando gli elementi definiti nel Content Standards for Digital Geospatial Metadata i soggetti interessati (governativi, non-profit o commerciali) possono rendere i loro cataloghi di informazioni spaziali, attraverso Internet, accessibili a tutti e oggetto di ricerche. »
(http://www.fgdc.gov/clearinghouse/clearinghouse.html) (ultimo accesso 14/10/2000)

«I tipi di ricerca possono essere basati sulla localizzazione, riguardare un certo periodo di tempo, essere ricerche libere sul testo oppure ricerche per campi, riguardare uno o più cataloghi»
(http://clearinghouse1.fgdc.gov/FGDCgateway.html) (ultimo accesso 14/10/2000)

«La " Geospatial Data Clearinghouse" è un insieme di più di 100 server di dati spaziali…»
(http://130.11.52.184/) (ultimo accesso 14/10/2000)

[4] Dati geospaziali, spaziali, geografici: qual è la differenza? Secondo alcuni la dizione "dati geografici" riguarderebbe i dati che si riferiscono a oggetti tradizionalmente rappresentati alle scale geografiche; sarebbe quindi una dizione limitativa rispetto a "dati spaziali". In Europa la dizione più usata, finora, è "dati geografici".

2.3 I metadata

Attraverso la Clearinghouse sono, in primo luogo, accessibili i *metadata*: *«I metadata, o "dati sui dati" descrivono contenuto, qualità, condizioni e altre caratteristiche dei dati* (FGDC, 1994).

I *metadata* servono a:

- *«conservare l'investimento interno in dati geospaziali che una organizzazione ha fatto;*
- *fornire informazioni sui dati disponibili presso una organizzazione ai cataloghi di dati, alle* clearinghouses *e ai brokers e*
- *fornire le informazioni necessarie per il trattamento e l'interpretazione dei dati ricevuti attraverso trasferimento da una fonte esterna.»*

(FGDC 1994, p. V)

Curiosamente (ma forse non tanto) la prima ragione data per documentare bene i dati è interna, quella di conservare l'investimento fatto in dati geografici: è certo che se i dati non sono documentati, con l'andar del tempo o a causa dei cambiamenti del personale, si può perdere la conoscenza del contenuto o della qualità dei dati.

Il Federal Geographic Data Committee ha sviluppato, attraverso processi partecipativi, un proprio standard riguardante il contenuto dei *metadata*, articolato in 7 sezioni, che sono:

- *Identification Information:* contiene titolo, area geografica coperta, un abstract e spesso informazioni aggiuntive che descrivono in modo libero i criteri con cui il *data set* è stato costruito, eventuali problemi sottesi ecc.;
- *Data Quality Information:* contiene una valutazione della qualità dei *data set* (ad es. l'accuratezza cartografica);
- *Spatial Data Organization Information*: descrive il modo in cui sono rappresentate spazialmente i dati, direttamente (dati raster o vector) o indirettamente (attraverso un indirizzo o un codice);
- *Spatial Reference Information*: descrive il tipo di proiezione geografica e il sistema di coordinate usate;
- *Entity and Attribute Information*: contiene, per ogni tipo di entità geografica, la descrizione dei suoi attributi; più database possono essere riferiti allo stesso tipo di entità; sono riportate anche le spiegazioni necessarie per decodificare i diversi valori che un attributo può assumere;
- *Distribution Information:* contiene le informazioni, anche quelle banali, per ottenere il *data set*;
- *Metadata Reference Information:* contiene la "firma" di chi ha compilato i *metadata*.

2.4 Il Framework

«Le applicazioni GIS nei diversi campi hanno la necessità ricorrente di un numero limitato di tipi di dati. Il framework è uno sforzo di tipo collaborativo e comunitario attraverso il quale questi dati comunemente necessari sono prodotti, mantenuti e integrati da parte di organismi pubblici e privati per una determinata area. Le organizzazioni locali, regionali, statali e federali e le imprese private vedono il framework come un modo per condividere risorse, migliorare l'interscambio e incrementare l'efficienza. »
http://www.fgdc.gov/framework/framework.html (ultimo accesso 18/10/2000)

«Il framework …fornisce i temi e i dati geografici più comunemente utilizzati e nello stesso tempo un ambiente che supporti lo sviluppo e l'utilizzo dei dati medesimi. Gli aspetti chiave del framework sono:

- *sette temi di dati geografici numerici comunemente usati*[5]*;*
- *procedure, tecnologie e linee di indirizzo necessarie ai fini dell'integrazione della condivisione e dell'uso di tali data; e*
- *il quadro istituzionali e le pratiche commerciali che incoraggino il mantenimento e l'uso di tali dati. »*

«*Integrando i propri dati geografici, che possono essere relativi a innumerevoli temi, con i dati comuni del framework, gli utilizzatori possono costruire le loro applicazioni più facilmente: I sette temi forniscono:*
- *i dati di base che possono essere usati nelle applicazioni;*
- *una base alla quale gli utilizzatori possono aggiungere o relazionare i dati geografici di dettaglio o attributi alfanumerici;*
- *gli elementi di riferimento* [punti fissi di riferimento geodetico]*su cui appoggiare le costruzione o la georeferenziazione dei propri data set;*
- *basi cartografiche per la rappresentazione delle localizzazioni o dei risultati di un'analisi di altri dati. »*

http://www.fgdc.gov/framework/overview.html (ultimo accesso 18/10/2000)

Il *framework* è la risposta a due esigenze abbastanza diverse, che nei documenti non vengono ben distinte. I dati cartografici da rendere disponibili possono essere di base perché sono comunemente usati in relazione a scopi diversi (come ad esempio una base cartografica per *census tract*) oppure perché di riferimento per l'organizzazione di altri dati (punti fissi di riferimento geodetico).
A seconda dei casi, gli utilizzatori dei dati avranno competenze diverse, in particolare gli utilizzatori del secondo tipo di dati dovranno avere elevate competenze specialistiche. Va rilevato che il *framework*, e più in generale la NSDI, non è rivolta agli utenti generici dell'informazione geografica, ma a utilizzatori professionali.

La costruzione del *framework* (secondo anche alcuni attori del processo, che hanno scritto in quegli anni, intorno al 1994) è fortemente legata a due principi che sono i capisaldi della politica federale USA per quanto riguarda l'accesso all'informazione: l'accesso gratuito ai dati (al puro costo di riproduzione) e l'assenza di copyright:

«Il concetto di costituire un framework di dati geospaziali derivati da una varietà di fonti dipende dalla possibilità di accedere liberamente e facilmente integrare i dati. Le politiche che in qualche modo rendono i dati indisponibili possono scoraggiare l'uso dei dati più appropriati per una determinata area. »
(Tosta, 1994).

[5] I sette temi sono: punti fissi di riferimento geodetico, ortoimmagini digitali, modello digitale del terreno, infrastrutture di trasporto, idrografia, limiti amministrativi, catasto.

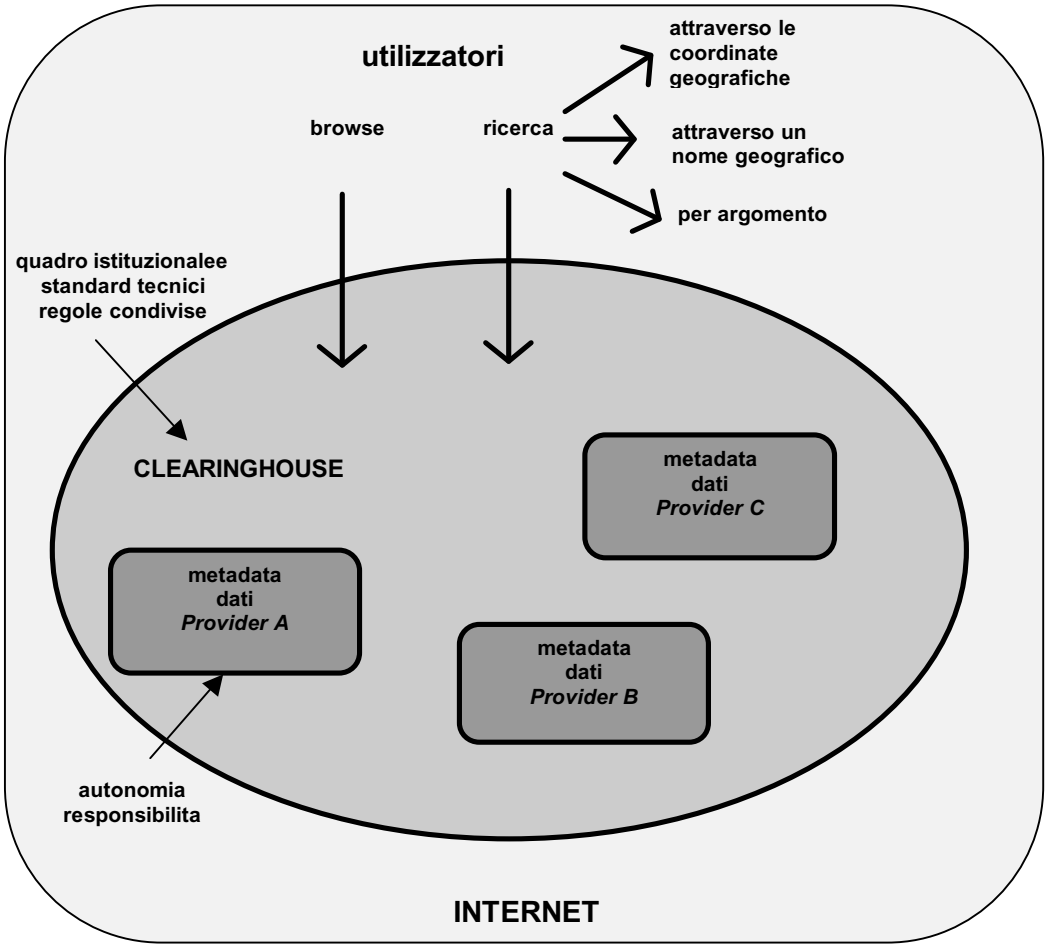


Figura 2. Schema concettuale della National Spatial Data Infrastructure

2.5 Una strategie partecipativa

La strategie di costruzione della NSDI comprende l'organizzazione di momenti di informazione e di dibattito e lo sviluppo di programmi di formazione e di addestramento.
Anche gli standard sono stati messi a punto attraverso la partecipazione, in particolare quello riguardante il contenuto dei *metadata*.
Forte enfasi viene anche posta sulla costruzione di compartecipazioni finanziarie. Tali *partnership* si basano sul riconoscimento da parte dei singoli soggetti dell'interesse, rispetto al raggiungimento dei propri scopi istituzionali (pubblici o privati), per la condivisione dell'informazione geografica, che ha alto costo.

2.6 Gli elementi di contesto: Internet

Gli elementi di contesto sono: Internet, le volontà politiche, il quadro normativo, i provvedimenti amministrativi e, su un piano diverso, l'Open GIS Consortium.
Quando il progetto della NSDI è stato avviato, nel 1990, Internet non era ancora, neppure negli Stati Uniti, il fenomeno esplosivo di questi ultimi anni. Nel 1994, quando Clinton firma l'Executive Order, il WWW (*World Wide Web*) stava appena cominciando a diffondersi.
Non c'è dubbio che ci sia un nesso forte tra Internet, la tecnologia WWW e la NSDI:
- le informazioni dislocate in siti diversi sono viste come un tutt'uno, Internet è il "luogo" dove

sta la Clearinghouse, che possiamo vedere come la piazza virtuale nella quale domanda di informazione e offerta si confrontano;
- la struttura ipertestuale permette di vedere diversi strati di informazioni (dello stesso tipo o di tipo diverso) interconnessi;
- Internet è anche lo strumento di accesso ai dati, dà la possibilità di prelevare quello che serve (files di dati, e anchesw ecc.).

Nel progetto della NSDI, Internet non è solo lo strumento: l'organizzazione "debole" di Internet è il paradigma della NSDI[6].
I diversi soggetti che rendono disponibili i cataloghi dei loro dati attraverso la Clearinghouse, mantengono piena autonomia sulle proprie scelte: decidono quali dati produrre, quali condividere, e hanno piena responsabilità della qualità dei dati.
Il coordinamento (la standardizzazione) riguarda pochi elementi, quelli indispensabili per l'accesso ai dati e per l'interscambio dei dati.
Per altri aspetti non c'è che da assumere gli standard di Internet.
La NSDI mi pare una puntuale applicazione del principio di sussidiarietà, anche se in nessun documento tra quelli che ho letto il termine è riportato.

Per quanto riguarda la questione dei formati informatici dei dati geografici, la NSDI non ha promosso quasi nessuna specifica iniziativa di standardizzazione e ha affidato la possibilità di condivisione e interscambio di dati geografici ai processi di standardizzazione di fatto o ai processi che l'OpenGIS Consortium promuove.
L'OGC è un consorzio sorto intorno al 1995 nel quale sono presenti oltre ai produttori di sw, produttori di dati, università, utilizzatori in genere.
La questione dei formati informatici dei dati geografici è complessa perché complessi sono i dati geografici, che devono trattare forma, posizione e relazioni degli oggetti sul territorio. Inoltre il significato dei dati geografici è diverso a seconda dell'ambito e degli scopi per cui i dati sono stati prodotti.
La storia del GIS è di una molteplicità di formati e di modelli logici dei dati: i formati sono quasi sempre proprietari, cioè coperti da *copyright*, legati a specifici sw e non esplicitamente documentati.
La nascita del OpenGIS Consortium e la sua crescita corrisponde al fatto che i produttori di sw hanno capito, finalmente, che i formati proprietari, più che una protezione delle proprie aree di mercato, sono uno sbarramento allo sviluppo complessivo del mercato dell'informazione geografica, che è di gran lunga più interessante (e remunerativo).
Il progetto principale che l'OGC sta portando avanti riguarda appunto il problema dell'integrazione fra dati geografici nati in ambienti diversi e gestiti dai sistemi diversi, e ha come obiettivo l'Open Geodata Interoperability Specification (OGIS), la definizione delle specifiche che devono permettere l'uso dei dati in ambienti diversi senza la loro traduzione, cioè appunto l'interoperabilità.
Questo obiettivo sta diventando più importante proprio con il diffondersi di Internet e delle applicazioni webGIS.

[6] Parlando di organizzazione "debole" di Internet mi riferisco più agli anni dell'impianto della NSDI (meta anni '90) che a oggi: Internet sta diventando sempre meno una struttura aperta e con poche regole, sempre di più una realtà da "consumare".

2.6 Gli elementi di contesto: il quadro giuridico

Certamente la NSDI ha beneficiato di una tradizione di condivisione di dati, assai più radicata in Usa che in Europa.

Ha beneficiato anche di un costante e deciso supporto politico ai massimi livelli che ha visto impegnati il vice presidente Al Gore, per tutti gli otto anni del suo mandato, dal 1993 al 2000, e a più riprese il Segretario degli Interni (Masser, 1998, p. 89).

Vista da questa parte dell'Atlantico, la NSDI, ancorché ancora un realtà *in progress*, appare un successo. E' utile capire quali sono i fattori chiave che lo spiegano.

Due fattori hanno avuto un ruolo decisivo nel decollo della NSDI:

- l'ampia disponibilità di dati geografici a costo zero,
- le norme relative al *copyright.*

Due leggi regolano la materia a livello federale negli USA.

Il Freedom of Information Act (1966) stabilisce che il costo dei dati federali deve essere definito ad un livello tale da ripagare il costi della loro duplicazione e distribuzione ma non più alto. Una Circolare del Office of Management and Budget del 1993 (OMB 1993) afferma che *"Il libero flusso di informazioni tra governo e cittadini è essenziale per una società democratica"* e che è protetto *"il diritto di accesso dei cittadini alle informazioni governative"*. Nel 1996 il FOIA è stato emendato con l'Electronic Freedom of Information Act che garantisce il pubblico accesso, elettronico, all'informazione amministrativa federale.

La seconda questione è definita dal Copyright Act del 1976, che specifica che non esiste *copyright* per lavori prodotti da impiegati o funzionari dello governo federale come parte dei loro specifici compiti.

Questo vuol dire anche che il settore privato può confezionare un prodotto basato sull'informazione federale, aggiungendovi valore, e rivenderlo senza che questo possa essere limitato dall'ente federale che ha prodotto i dati.

Libero accesso e *copyright* sono interconnessi: il dibattito in Europa su questi temi è molto aperto (complicato ovviamente dal fatto di dover fare i conti con leggi e norme diverse da paese a paese). Bisogna riflettere sul fatto che in questo dibattito non c'è spazio per enunciazioni di principi disconnesse dalla pratica. Se il principio del diritto di accesso dei cittadini alle informazioni governative, sancito dalle leggi federali USA, è applicato all'informazione geografica, bisogna almeno riflettere sul fatto che i cittadini non sono assolutamente tutti eguali: alcuni non hanno, di fatto, la capacità tecnica di accedere a questa informazione, altri non solo hanno la capacità di accedervi ma anche di farne la fondamentale materia prima per i loro affari.

Le norme giuridiche citate valgono a livello federale. A livello degli stati e delle contee i comportamenti sono vari. Secondo una *survey* del 1996 riguardante tutti gli USA (citata da Masser, 1998, p. 81) si va da una politica di libero accesso (che vuol dire: dati a costo zero o al puro costo del supporto, costo complessivo per l'intero database minore di 1.000 $, nessuna restrizione per quanto riguarda gli usi secondari) a politiche di mercato (in cui i 3 criteri citati vengono disattesi) passando per soluzioni intermedie (in cui 1 o 2 dei criteri citati vengono disattesi): le soluzioni intermedie sono circa il 56%, quelle "libero accesso" il 24%, quelle basate sul prezzo di mercato il 20%.

Non c'è dubbio che però il Freedom of Information Act e il Copyright Act, norme radicate nella prassi federale da circa 30 anni, hanno avuto una funzione fondamentale rispetto alla costruzione della NSDI, oltre che per i motivi riportati al punto 2.4, anche perché in questi 30 anni il facile accesso all'informazione geografica si è legato, in un circolo virtuoso, con lo sviluppo dei web e di

applicazioni GIS, oltreché con la crescita degli utilizzatori di informazione geografica.

2.6 Il ruolo svolto dal Bureau of the Census

Il Freedom of Information Act e il Copyright Act hanno influenzato in modo preciso il comportamento di soggetti quali il Bureau of the Census (il corrispondente americano dell'ISTAT). Il Bureau of the Census già per i censimenti del 1970 e 1980 aveva costruito, per le sole aree più urbanizzate, una base cartografica a grande scala, strutturata topologicamente per la gestione e la rappresentazione dei dati censuari per *census tract* e per isolati (DIME files) (Huxhold, 1991, pp. 163-171). Per il censimento del 1990 questa base è stata estesa all'intero territorio degli USA, rigorosamente georeferenziata, e rivista per quanto riguarda il modello concettuale dei dati. Il costo dell'intera operazione è stato certamente molto elevato ma il risultato è stato di disporre di una base cartografica a grande scala, estesa a tutto il territorio nazionale, consistentemente strutturata secondo il modello TIGER (Topologically Integreted Geographic Encodind and Referencing).

Le libera disponibilità delle basi DIME prima, e poi di quelle TIGER, spiega, in buona misura, il divario esistente tra USA e Europa per quanto riguarda la diffusione dell'utilizzo dell'informazione geografica.
Cito ad esempio la possibilità di inserire nelle proprie pagine web una mappa interattiva con la rappresentazione geografica dei propri dati, anche senza disporre delle necessarie basi cartografiche e della tecnologia webGIS, ma utilizzando il servizio di "generazione remota di mappe" che Bureau of the Census mette a disposizione, servizio basato appunto sulle basi TIGER (http://tiger.census.gov) (ultimo accesso 8/9/2000).

L'azione del Bureau of the Census e la NSDI, che migliora l'accesso ai dati primari e crea una abitudine all'uso della rete, induce probabilmente anche altre iniziative di distribuzione "secondaria" di dati geografici attraverso la rete.
Un esempio è il Demographic Data Viewer, messo a punto dal Consortium for International Earth Science Information Network (Ciesin)[7], accessibile al sito (http://plue.sedac.ciesin.org/plue/ddviewer) (ultimo accesso 7/9/2000).
Ma mentre il NSDI si rivolge ai professionisti dell'elaborazione di dati geografici, il DDviewer si rivolge a utenti finali.

3. La European Geographic Information Infrastructure

L'Executive Order di Clinton suscitò un notevole interesse in Europa e anche un po' di emozione nel mondo GIS europeo. Già nel 1994 i ministri di Francia, Germania, Spagna e Olanda avevano richiesto alla European Commission di prendere una iniziativa politica per la creazione di un database di informazione geografica continuo e omogeneo relativo all'intera Europa
Il contesto è l'enfasi posta in generale sul ruolo dell'informazione[8]. Sull'onda della NSDI matura la proposta di definire e implementare una European Geographic Information Infrastructure (EGII).

---

[7] La CIESIN è un consorzio fondato nel 1989 come ente di diritto privato, non a scopo di lucro, i cui membri sono alcune importanti università e enti non governativi di ricerca. Il suo obiettivo è *«incoraggiare lo studio interdisciplinare del cambiamento ambientale globale»* rendendo accessibili e integrando dati fisici, naturali e socioeconomici trasversali rispetto a diverse area disciplinari.
[8] Vedi il molto citato libro bianco di Delors del dicembre 1993 (CEC, 1993) che individua nello sviluppo di una società dell'informazione la sfida del 21° secolo per l'Europa.

Nel 1995, la DG XIII (Directorate General for Telecommunications, Information Market and Exploitation of Research) organizza una consultazione sull'idea della EGII e viene prodotto una prima versione del documento *GI2000. Toward a European Policy Framework for Geographic Information*. Gli obiettivi erano di identificare le barriere allo sviluppo di infrastrutture di dati geografici e di considerare le possibili soluzioni

> *«La EGII vuole essere un insieme stabile di regole, standard e procedure, definite a livello europeo, per la creazione, raccolta, scambio e uso di informazioni geografiche (GI). La EGII vuole anche assicurare che i data set di base siano immediatamente disponibili a livello europeo e che esista un servizio di i data set di base in modo che tali dati possano essere facilmente individuati dai potenziali utilizzatori.»*
> (http://www2.echo.lu/gi/gi2000/en/egii/sum.html) (ultimo accesso 16/6/1996)

Gli ingredienti sono quindi gli stessi della NSDI: un meccanismo scambio di informazioni geografiche, i *medatada*, i *data set* di base, un quadro di regole e standard, e ovviamente Internet. Le questioni ritenute più rilevanti, da approfondire, sono:
- gli aspetti legali (salvaguardia della privacy, copyright dei dati, aspetti legali connessi alla qualità dei dati ecc.);
- il mercato di dati (le distorsioni generate dalle diverse politiche di accessibilità ai dati praticate nei diversi paesi, dalla presenza di dati di pubblico domino, le politiche dei prezzi ecc.).

Il documento *GI2000* è oggetto di diverse revisioni; circolano via e-mail versioni emendate secondo diverse visioni nazionali; si svolgono *meetings* (l'ultimo è del marzo 1999)... Ma non viene raggiunto l'obiettivo previsto che era l'adozione del documento da parte della European Commission, poi del Consiglio dei Ministri e del Parlamento Europeo.

Un motivo del non raggiungimento dell'obiettivo previsto viene individuato nel mancato supporto dell'industria (Annoni, 2000 p. CIV). In Europa la *"GI Industry",* più che nei produttori di sw, un settore nel quale l'Europa è pressoché totalmente assente, va identificata nel settore che produce e usa l'informazione geografica (Longhorn, 1999). Ma qual è la dimensione del mercato per l'informazione geografica?
A mio avviso il non risultato raggiunto è da mettere in relazione con l'incapacità-impossibilità di attivare azioni concrete (a livello europeo, ma necessariamente anche a livello dei singoli paesi) a favore della condivisione dell'informazione geografica, per un effettivo più facile accesso ai dati e un loro più diffuso uso. In Europa tutto ciò sembra bloccato dalle preoccupazioni relativa al mercato e alla proprietà dei dati, e istituzionalmente mancano soggetti che possano svolgere il ruolo che il Bureau of the Census ha svolto in USA, di creazione della massa critica di dati geografici in grado di innescare il processo di diffusione del loro uso.

Va ricordato che il grosso dell'informazione geografica è pubblica, e che l'80-90 % dell'informazione pubblica è geografica, cioè contiene elementi georefenziabili (o direttamente attraverso coordinate, o indirettamente attraverso indirizzo, un cap ecc.).
Questo spiega perché l'attenzione della *GI Community* europea nel 1998-1999 si focalizzata intorno sul *Libro verde sull'informazione del settore pubblico nella società dell'informazione* (Commissione Europea, 1998) della fine del 1998. Così su 180 reazioni ricevute su questo libro verde, non indirizzato al tema della informazione geografica 40 provengono dalla comunità dell'informazione geografica (Annoni, 2000, CIV).

Rispetto alla prospettiva della costruzione della EGII, l'iniziativa in corso che in corso che sembra diventata di riferimento è il progettoETeMII (Accompanying measure to support the setting up of a European Territorial Management Information Infrastructure), nell'ambito del V programma quadro, operativo dal gennaio 2000. Il suo obiettivo è aiutare a sviluppare gli elementi chiave dell'Infrastruttura Geografica; quelli richiamati sono ben noti: i dati di riferimento, le politiche di accesso ai dati, l'interoperabilità e gli standards, i *metadati* (http://www.ec-gis.org/etemii) (ultimo accesso 24/10/2000).
Gli elementi nuovi mi sembrano due:
- uno sforzo per creare consenso, coinvolgendo gli attori attivi nel settore a livello europeo e a livello dei singoli paesi;
- una maggiore concretezza nell'individuare i terreni su cui lavorare.

Uno di questi terreni, come detto, è la costruzione dei "dati di riferimento":

> *«l'obiettivo e quello di raggiungere un consenso sul piano tecnico…riguardo ad una definizione comune dei dati di riferimento, e di rendere tali dati disponibili ad un prezzo sostenibile. In specifico, l'obiettivo è quello di rispondendo alla domanda dei maggiori utenti che chiedono dati da utilizzare a scala locale* [alla grande scala] *e delle attività a valore aggiunto sui dati geografici, con lo sviluppo anche di nuove forme di cooperazione tra soggetti pubblici e privati. »*
> (ETeMII, 2000)

La definizione dei dati geografici essenziali dovrà avvenire in base alle esigenze degli utenti rispetto a due nuovi concetti,*e-commerce* e *e-government,* emblematici di due approccio, uno di tipo più privato l'altro pubblico, all'uso di Internet.
Questa stessa concretezza si manifesta anche sugli altri terreni, ad esempio per quanto riguarda i *metadata* si sottolinea che bisogna partire dai bisogni degli utenti, raggiungere un buon livello di consenso, realizzare seminari di lavoro … La specificità europea emerge quando si dichiara la necessità di porre attenzione sugli aspettimulti-linguistici e multi-culturali che caratterizzano l'Europa.

4. La Global Spatial Data Infrastructure

Per completare il quadro, accenno all'iniziativa della GSDI, che è cresciuta fondamentalmente attraverso l'organizzata di conferenze annuali, la prima tenuta nel 1996. Uno dei risultati è stato un manuale, lo *SDI Implementation Coookbook*, scaricabile dal sito www.gsdi.org (Nebert, 2000).
In esso si ritrovano i concetti base già visti.

> *«Il termine "Spatial data Infrastrucure" è spesso usato per denotare una rilevante insieme di tecnologie, politiche, atti istituzionali che facilitano la disponibilità e l'accesso a dati spaziali. La SDI fornisce le basi per l'individuazione dei dati spaziali, la loro valutazione, e il loro uso da parte di utilizzatori o providers, nei diversi livelli di governo, nel settore commerciale, nel settore non-profit, nelle università a da parte dei cittadini in genere.»*

La parola infrastruttura è usata per sottolineare il concetto di affidabilità e quello di contesto di supporto che deve facilitare l'accesso all'informazione geografica.
I componenti della GSDI richiamati nel*SDI Implementation Coookbook* sono:
- i *framework* o core data;
- i *metadati,* visti ovviamente come strumento base per la ricerca, la valutazione e l'uso dei dati geografici, da produrre con riferimento agli standard emergenti e da implementare usando le soluzioni freeware o sw a basso costo;
- il catalogo dei dati.

Un componente della SDI è il webGIS (qui chiamato *on-line mapping*), che risponde alle

necessità di informazione geografica da parte di utilizzatori non specialisti o di soggetti interessati alla consultazione dei dati più che al loro uso (cioè che non hanno bisogno del *download* dei dati).
Trovati i dati geografici di proprio interesse attraverso il webGIS o attraverso il catalogo, l'ultimo componente della SDI riguarda ovviamente l'accesso ai dati.
Altre all'introduzione del webGIS, un secondo elemento innovativo dell'approccio della GSDI è la distinzione fra diversi tipi di utilizzatori (non presente nella NSDI), i *novice or browse users* e gli *advanced users*

6. Italia il Sistema di comunicazione dati territoriali AIPA
L'unica esperienza italiana è il progetto intersettoriale Sistemi Informativi Territoriali, che avviato dall'AIPA nel 1996, è tuttora in una fase di studio
Il progetto riguarda fondamentalmente l'interscambio di dati tra pubbliche amministrazioni, e solo in un fase successiva tempo con soggetti privati.
I componenti sono quelli già visti: manca però la componente *framework data.*
Mi ha sorpreso il fatto che nei documenti pubblicati su Internet non ci siano riferimenti né al dibattito europeo nè all'esperienza americana. Gli unici richiami riguardano la normativa tecnica europea sull'informazione territoriale (CEN-TC 287) e quella internazionale (ISO 211, FDGC e OpenGIS).
Trattandosi per ora di scambi tra pubbliche amministrazioni manca ovviamente qualsiasi approfondimento sui temi connessi al prezzo dei dati e al *copyright*.

6. Conclusioni: mettendo insieme webGIS e infrastrutture di dati geografici
Jack Dangermond, il patron di ESRI, nel 1998 parla di transizione dal (1) *personal GIS*, nel quale i singoli accedono ai propri dati sulla proprie macchine; al (2) *departmental GIS,* nel quale diversi utilizzatori di un dipartimento (ad esempio di un ente locale o di un ministero) accedono ai data base dipartimentali; al (3) *corporate GIS*, nel quale molti utilizzatori di un ente (ad es. un comune o una regione, organizzato su più dipartimenti o settori) accedono ai database gestiti centralmente; al (4) *societal or internet GIS*, nel quale molti utilizzatori accedono a molti database localizzati ovunque (riportato da Heikkila, 1998, p. 354).
Il titolo dell'articolo di Heikkila è "GIS is dead, long live GIS", che evoca il grido del banditore medioevale che annunciava in contemporanea la morte del re e l'ascesa al trono di un nuovo re, sintetizza una tesi di fondo del tutto condivisibile: Internet sta cambiando radicalmente il GIS.

Il rapporto molti-a-molti del *societal or internet GIS* diventa meno vago se mettiamo insieme l'evoluzione della tecnologia GIS nella direzione webGIS con l'idea della infrastruttura di dati geografici.
WebGIS[9] significa i servivi e le tecniche per accedere a dati geografici attraverso Internet utilizzando la tipica achitettura *client-server* del www.
L'accesso può riguardare:
- la consultazione di mappe statiche,
- l'interrogazione dinamica di mappe,
- interrogazioni complesse e analisi spaziali *on-line* di dati geografici,
- accesso a dati GIS attraverso Internet.

[9] Plewe 1997, tuttora l'unico testo che tenta una sistemazione della materia webGIS, affianca a weGIS anche la dizione Distributed Geographic Information (DGI): questa dizione che di fatto non ha avuto diffusione nel contesto di questo paper è particolarmente significativa.

L'ultimo tipo di accesso può avvenire attraverso:
- un *fat client*, cioè un *client* specialistico, un GIS viewer, (normalmente scaricabile gratuitamente dalla rete);
- un sw GIS completo capace di accedere ai dati utilizzando i protocolli di Internet, quello che Plewe chiamare *"net-savvy GIS*", cioè un GIS che conosce la rete.

L'approccio webGIS sta riguardando non solo l'accesso all'informazione geografica da parte di utenti generici ma anche la distribuzione dell'informazione geografica all'interno delle organizzazioni, a utilizzatori di diverso tipo (generici o specialistici), con diverse competenze, dotati di macchine con installatisw differenti.
Anche a questo livello, di singola organizzazione o ente, ritroviamo i concetti base e i componenti propri delle infrastrutture di dati geografici: il catalogo, *framework*, i *metadata,* gli standard.
L'infrastruttura di dati geografici può quindi essere vista come una idea proponibile a diversi livelli (aziendale, locale, nazionale, globale), livelli che possono in molti casi utilmente integrarsi.

"GIS is dead, long live GIS". Un GIS è morto per lasciare posto a qualcosa di contemporaneamente simile e diverso: molti soggetti; dati geografici diversi, dispersi nella rete; strumenti diversi, che interagiscono su livelli diversi (globali, locali); in un contesto di responsabilità diffuse e regole condivise, alcune su più livelli, altre livello per livello.
La dizione sistema informativo geografico può essere applicata a tutto ciò, ma bisognerebbe trovarne un'altra.

6. Riferimenti bibliografici

Annoni A., 2000, "Verso una infrastruttura dell'informazione geografica in Europa" in: ASITA, *Atti della IV Conferenza Nazionale "Informazione geografica: innovazione e formazione",* pp. CIII-CVIII.

Commission of the European Communities (CEC), 1993, **"**Growth, Competitiveness and Employment: the Challange and Ways Forward into the 21st Century", *Bulletin of the European Communities,* Supplement 6/93.

Commissione Europea, 1998, *L'informazione del settore pubblico: una risorsa fondamentale per l'Europa. Libro verde sull'informazione del settore pubblico nella società dell'informazione,* COM(98)585 def.

Craglia M, 1999, Public Sector Information: a Key Resource for Europe COM(1998)585. Response from the Association of Geographic Information Laboratories in Europe (AGILE). (http://castafiore.uni-muenster.de/agile/index.htm) (ultimo accesso 26/10/2000)

Craglia M., Annoni A., Masser I., eds, 2000, *Geographic Information Policies in Europe:National and Regional Perspectives. EUROGI-EC Data Policy Workshop. Proceedings, 15th November 1999* EUR 19522 EN.
(http://www.ec-gis.org/reports/policies.pdf) (ultimo accesso 10/10/2000)

Di Pace R, Nicoletti A., 2000, "Sistemi di controllo di flusso nella produzione e gestione di warehouses territoriali complessi" in: ASITA, *Atti della IV Conferenza Nazionale "Informazione geografica: innovazione e formazione",* pp. 741-742.

Eurogi, European Umbrella Organisation for Geographic Information, 2000, *Towards a strategy for geographic information in Europe. A consultation paper.*
(http://www.eurogi.org/geoinfo/eurogiprojects/strategy.pdf) (ultimo accesso 26/10/2000)

Executive Order 12906, 1994, Published in the April 13, 1994, edition of the Federal Register, Volume 59, Number 71, pp. 17671-17674.
(http://fgdc.er.usgs.gov/publications/documents/geninfo/execord.html) (ultimo accesso

16/10/2000)
FGDC (Federal Geographic Data Committee), 1994, *Content standard for digital geospatial metadata,* FGDC, Washington D.C.
FGDC (Federal Geographic Data Committee), 1998, *Content Standard for Digital Geospatial Metadata, (FGDC-STD-001-1998).*
Heikkila E.J., 1998, "GIS is Dead; Long Live GIS!", *APA Journal,* pp. 350-360.
Huxhold W.E., 1991, *An Introduction to Urban Geographic Information Systems,* New York, Oxford, Oxford University Press.
Longhorn R.A., 1999, *Geospatial Information / Spatial Information Infrastructure(s) – an Update*.
Masser I., 1998, *Government and Geographic Information*, Taylor & Francis, London.
Nebert D.D, ed, 2000, *Developing Spatial Data Infrastructure. The SDI Cookbook*, ver 1.0, 6 July 2000. (http://www.gsdi.org/cookbook) (ultimo accesso 18/10/2000)
Plewe B, 1997, *GIS on line: information retrival, mapping and the Internet*, Santa Fe, OnWord Press.
SAS, 1995, «The Data Warehouse», *Inform. The European SAS System Journal*, n 13, pp. 5-22.
Thewessen T., van de Velde R., 1995, "The Concept of a Data Warehouse; Methods and Experiences", in: *Join European Conference on Geographical Information 1995. Proceedings*, Basel, JEC, pp. 77.
Tosta N., 1994, "Data Policies and the National Spatial Data Infrastructure" in: NCGIA, Center for the Study of Law, Science and Technology Arizona State University, *Proceedings of the Conference on Law And Information Policy for Spatial Databases,* October 28-29, Tempe (http://www.spatial.maine.edu/tempe/tempe94.html) (ultimo accesso 18/6/1996).

Quinta
Conferenza
Nazionale
di Statistica

WORKSHOP

Connessione in rete: sicurezza informatica e riservatezza

coordinatore

Antonio Lioy
docente del Politecnico di Torino

Roma 15, 16, 17 novembre 2000
Palazzo dei Congressi
Piazzale Kennedy - EUR

ISTAT

WORKSHOP

Connessione in rete: sicurezza informatica e riservatezza

coordinatore

Antonio Lioy
docente del Politecnico di Torino

SOMMARIO

Prof. A. Lioy	Sicurezza delle reti e dei sistemi informativi
Dr. C. Giustozzi	Privacy e firma digitale
Ing. M. Terranova	La sicurezza nella Rete Unitaria
Prof. E. Bertino	Protezione della riservatezza in basi di dati
Ing. A. Guarino	Sicurezza informatica e produzione statistica

WORKSHOP

Connessione in rete: sicurezza informatica e riservatezza

Sicurezza delle reti e dei sistemi informativi

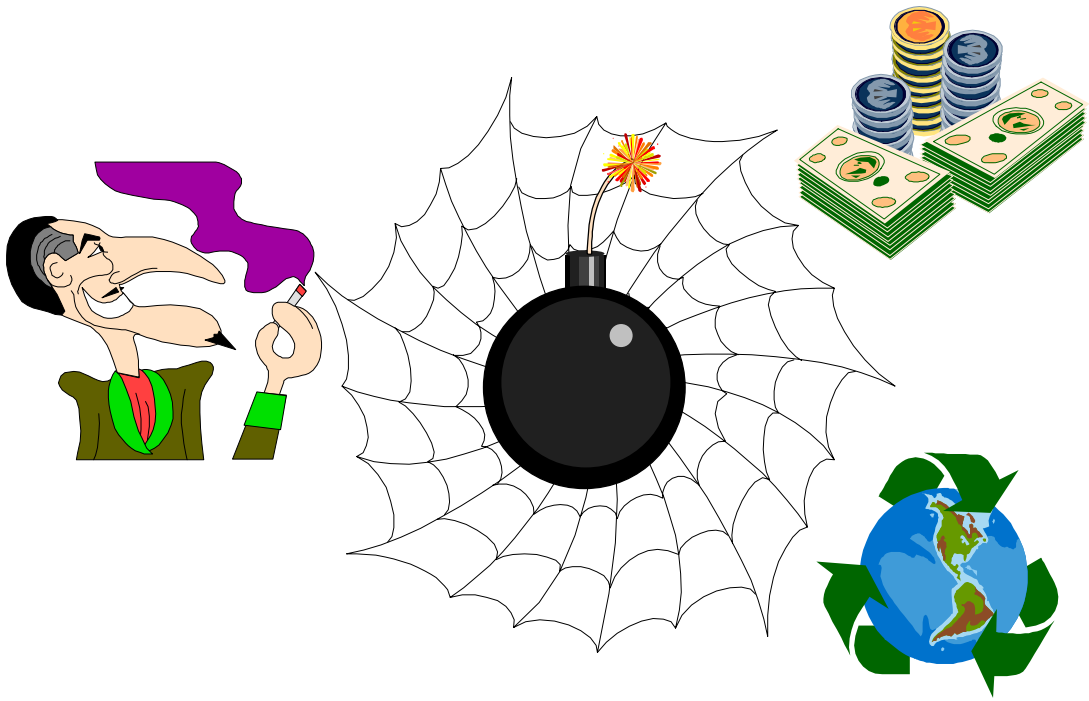
A. Lioy

Sicurezza delle reti e dei sistemi informativi

Antonio Lioy
< lioy @ polito.it >

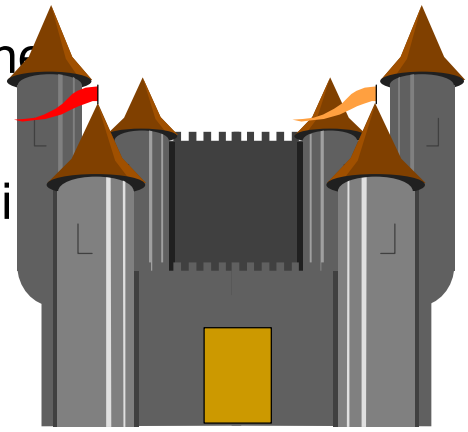
Politecnico di Torino
Dip. Automatica e Informatica

Perché è esploso il problema "sicurezza" ?



Sicurezza: dove è il nemico?

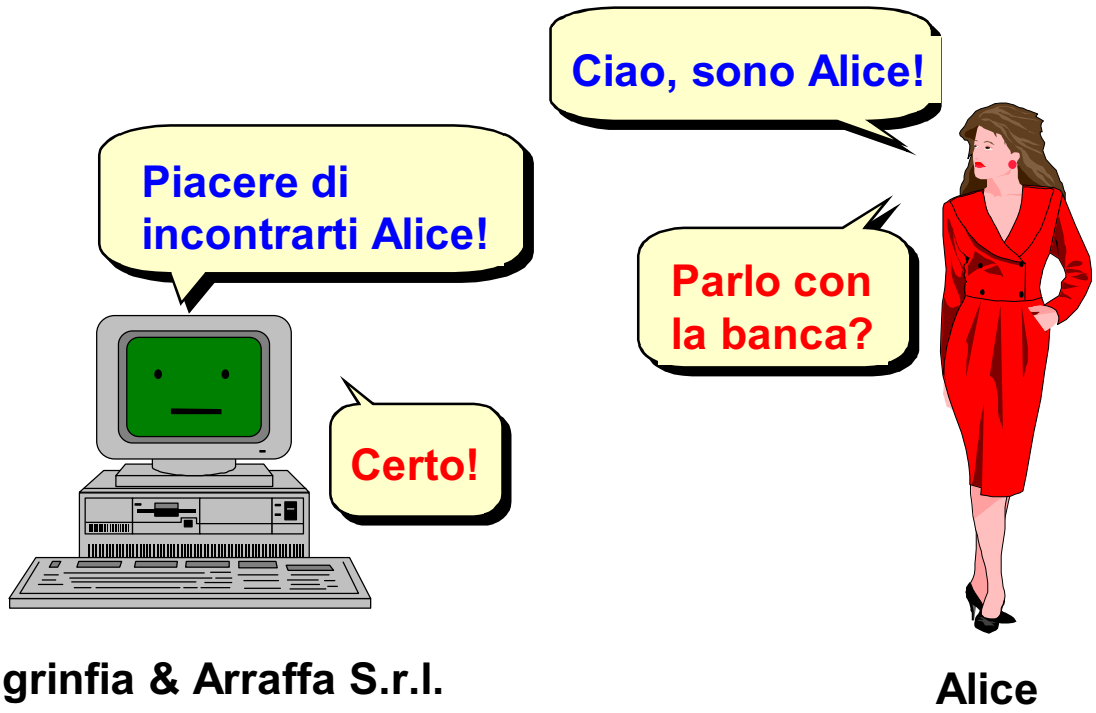
- **fuori dalla nostra organizzazione**
  - difesa del perimetro (firewall)
  - protezione dell'Extranet (VPN)
- **dentro la nostra organizzazione**
  - protezione della LAN / Intrane
- **ovunque**
  - protezione delle applicazioni



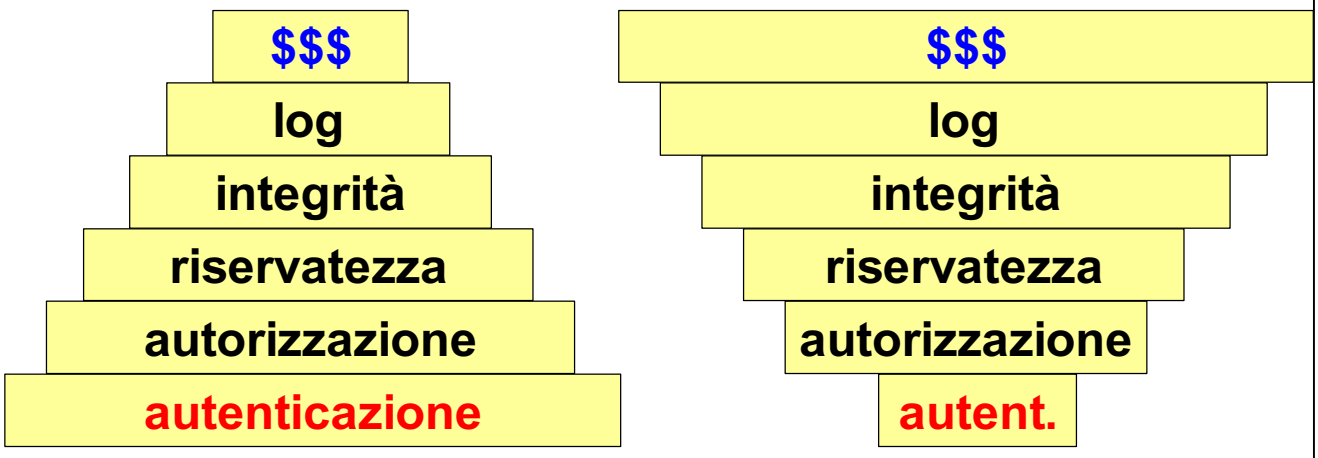
Proprietà di sicurezza

autenticazione semplice / mutua	***authentication simple / mutual***
autorizzazione	***authorization***
integrità	***integrity***
riservatezza	***privacy, secrecy***
non ripudio	***non repudiation***
disponibilità	***availability***
tracciabilità	***accountability***

Mutua autenticazione



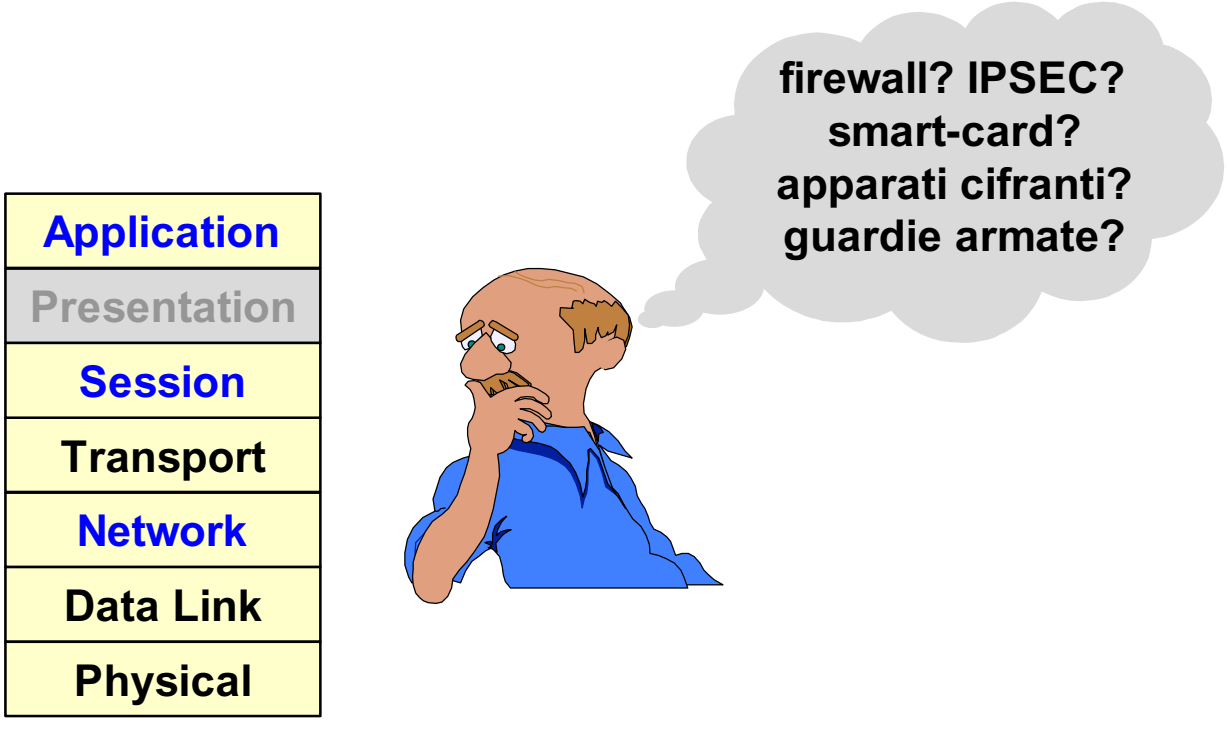
La piramide della sicurezza



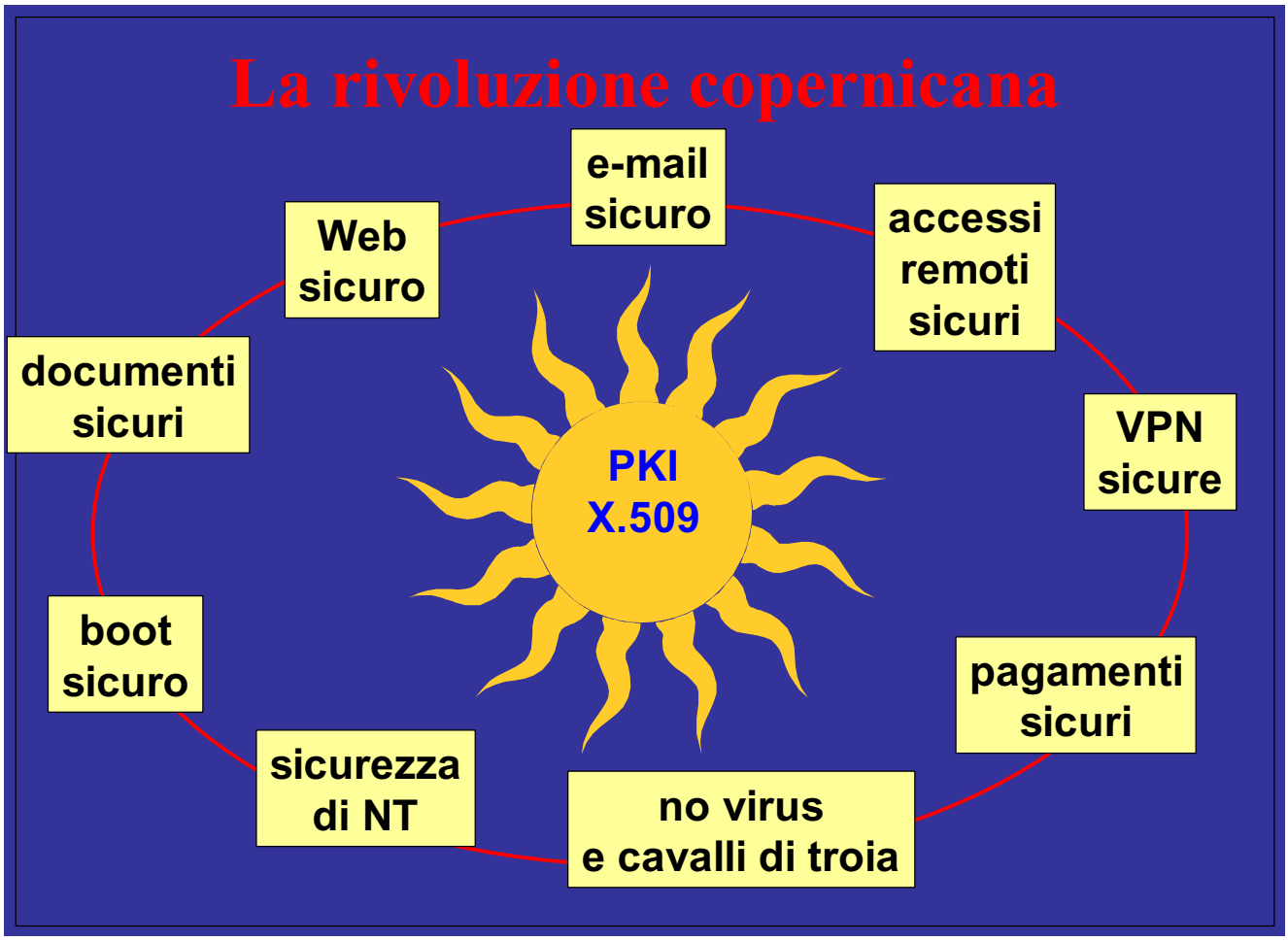
L'ABC degli attacchi

- **IP spoofing / shadow server**
  qualcuno si sostituisce ad un host
- **packet sniffing**
  si leggono password di accesso e/o dati riservati
- **connection hijacking / data spoofing**
  si inseriscono / modificano dati durante il loro transito in rete
- **denial-of-service**
  si impedisce il funzionamento di un servizio (es. la guerra dei ping)

A quale livello di rete è meglio realizzare la sicurezza?



La rivoluzione copernicana

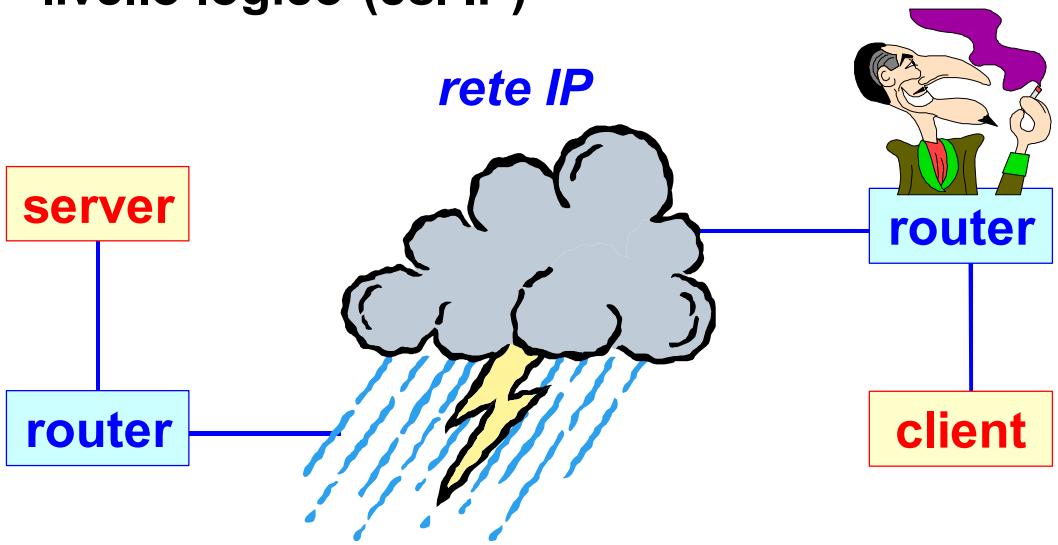


Esempio di certificato X.509

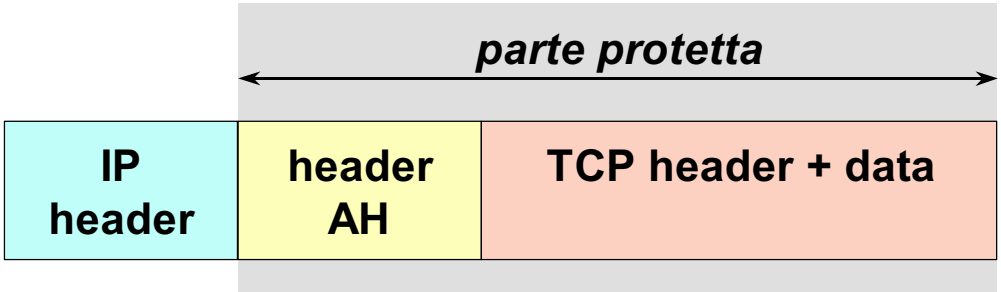
version	2
serial number	1231
signature algorithm	RSA with MD5, 1024
issuer	C=IT, O=Polito, OU=CA
validity	1/1/2000 - 31/12/2000
subject	C=IT, O=Polito, CN=Antonio Lioy, Email=lioy@polito.it
subjectpublickeyinfo	RSA, 1024, xx...x
CA digital signature	yy...y

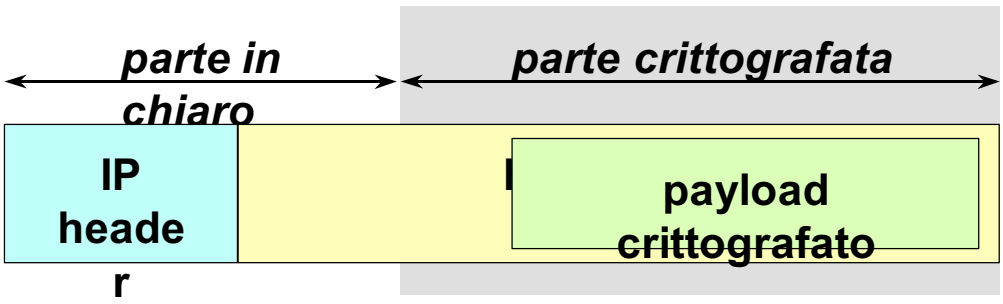
Sicurezza a livello network

- protezione end-to-end per reti omogenee a livello logico (es. IP)



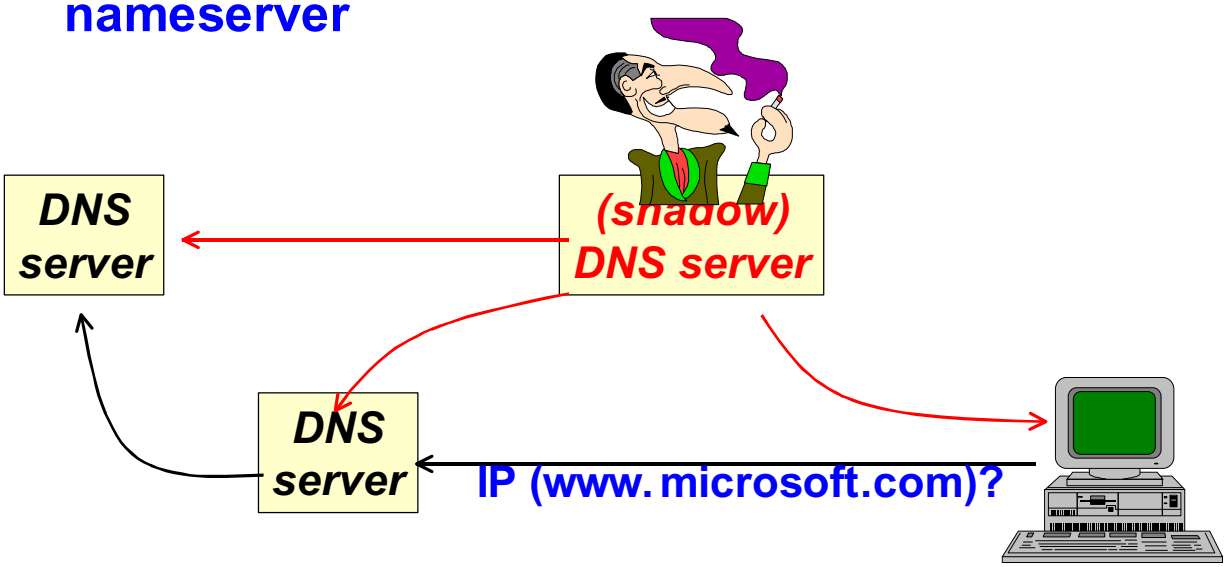
IPsec: formati AH e ESP





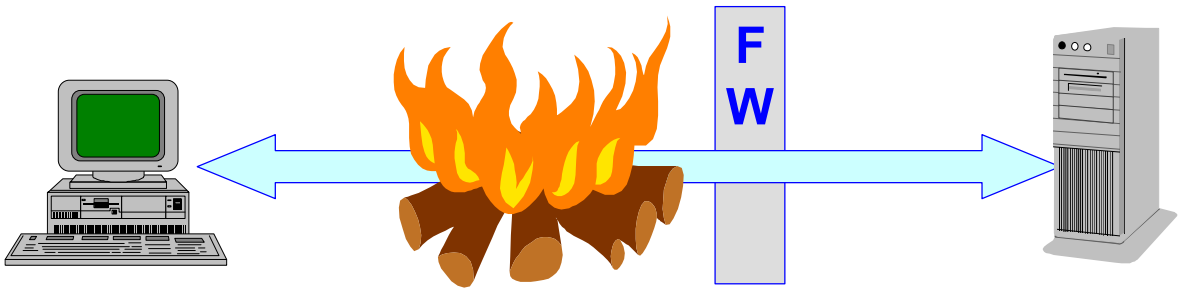
Sicurezza del DNS

- attacco 1: shadow server
- attacco 2: avvelenamento della cache
- DNS sicuro con chiavi RSA per ogni nameserver

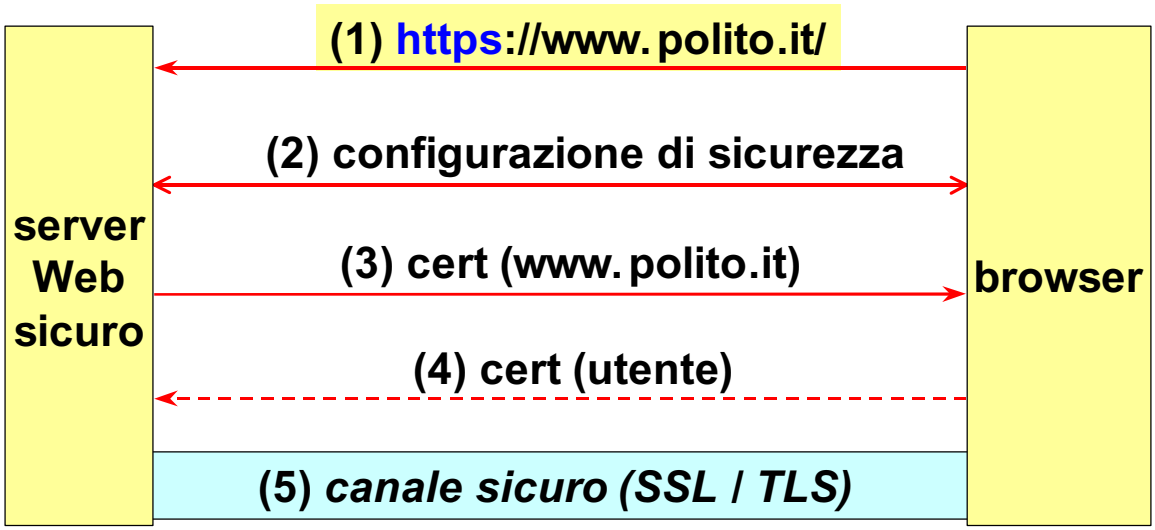


Firewall e sicurezza applicativa

- i firewall proteggono da attacchi sui canali che vengono chiusi
- … ma nulla possono fare per i canali che restano aperti
- per i canali aperti occorre protezione a livello applicativo



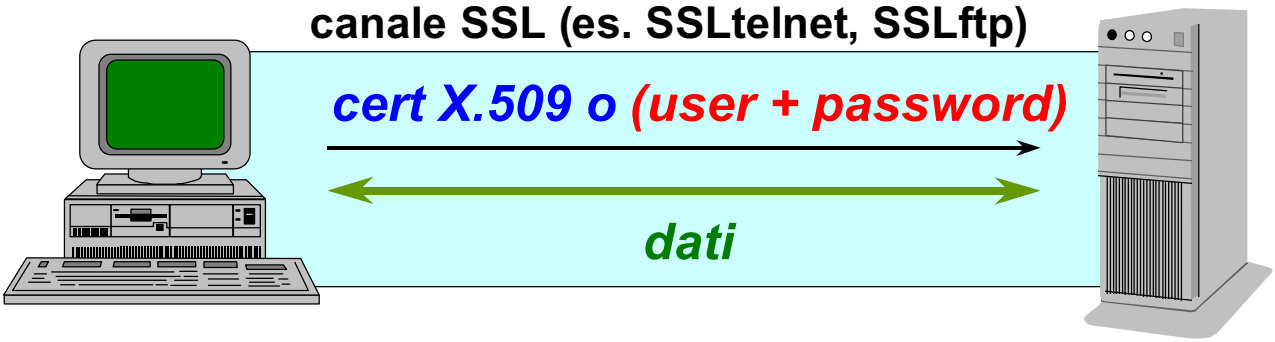
SSL / TLS



N.B. è sicurezza di canale, quindi non protegge i dati sul browser e sul server

Sicurezza dei collegamenti interattivi

- autenticazione del server
- autenticazione del client
- protezione del canale



Posta elettronica sicura: S/MIME

- utilizza firma digitale con certificati X.509v3
- protegge messaggi MIME

firmato

testo
tabella Excel
docum. Word
firma digitale in formato S/MIME

firmato e cifrato

testo
tabella Excel
docum. Word
firma digitale in formato S/MIME

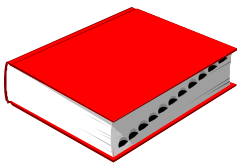
busta cifrata in formato S/MIME

cifrato

testo
tabella Excel
docum. Word

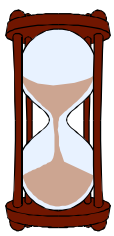
busta cifrata in formato S/MIME

Documenti elettronici



FORMATI STANDARD DEI:

- certificati
- firme digitali
- ruoli
- documenti
- marche temporali



WORKSHOP

Connessione in rete: sicurezza informatica e riservatezza

Privacy e firma digitale

C. Giustozzi

PRIVACY E FIRMA DIGITALE

CORRADO GIUSTOZZI
(C.GIUSTOZZI@IET.IT)

Le comunicazioni elettroniche

- Enorme potenzialità di sviluppo sociale:
  - comunicazioni fra privati
  - commercio elettronico
  - transazioni finanziarie
- Ma tutto dipende da alcune certezze:
  - identità dei corrispondenti
  - sicurezza della comunicazione
  - riservatezza della comunicazione
- La soluzione tecnologica è la *crittografia*

Esempio: la telefonia cellulare

- È difficile progettare un sistema sicuro:
  - contro situazioni accidentali
  - contro l'uso improprio o malizioso
- La rete di prima generazione (TACS):
  - non è protetta contro le intercettazioni
  - non è protetta contro l'impersonazione
- La rete di seconda generazione (GSM):
  - è basata su tecniche crittografiche
  - è sicura e protetta contro abusi e violazioni

La rete analogica TACS

- È stata progettata "con ingenuità":
  - supponendo la buona fede degli utenti
  - non considerando la possibilità di truffa
- Due gravi difetti di progetto:
  - trasmissione in chiaro della comunicazione
    - permette di intercettare le comunicazioni
    - consente la violazione della privacy degli utenti
  - trasmissione in chiaro dell'handshake/handover
    - permette di ricostruire l'identità del cellulare (clonazione)
    - consente le truffe ai danni del gestore della rete

La rete digitale GSM

- È stata progettata con attenzione:
  - utilizza tecniche crittografiche
  - previene truffe ed abusi
- La crittografia protegge a due livelli:
  - la comunicazione fra utenti
  - la comunicazione di servizio fra terminale e rete
- Il risultato è una rete sicura:
  - contro l'intercettazione delle comunicazioni
  - contro l'impersonazione e l'uso abusivo

La crittografia a chiave pubblica

Concetti di base e principi di funzionamento

Meccanismo di base

- Elementi di un sistema crittografico:
  - il “cifrario” (metodo o algoritmo)
  - la “chiave” (informazione)
- Si applica il *“Principio di Kerchoffs”*:
  - il metodo si deve supporre noto a tutti
  - il segreto risiede solo nella chiave
- La conoscenza della chiave:
  - permette di cifrare e decifrare documenti
  - può costituire prova certa di identità

Algoritmi a chiave *asimmetrica*

- Diffie ed Hellmann, 1976
- Il sistema usa *due* chiavi:
  - una chiave *diretta* K_d
  - una chiave *inversa* K_i
- Se K_d cifra, K_i decifra e viceversa
- La chiave che ha cifrato non può decifrare
- La conoscenza di una delle due chiavi non permette di ricavare l’altra chiave

Crittografia a chiave *pubblica*

- Nella PKC ogni utente ha *due* chiavi:
  - quella diretta viene resa *pubblica* (K_p)
  - quella inversa rimane *segreta* (K_s)
- Tutte le chiavi pubbliche sono in un elenco centralizzato, consultabile liberamente
- Ogni chiave segreta è nota solo al legittimo proprietario
- Ciascun utente genera le proprie chiavi

Proprietà dei sistemi PKC

- Gestione delle chiavi semplificata
- Si può scrivere a uno sconosciuto senza doversi preventivamente scambiare una chiave di cifratura in comune
- La propria chiave privata è e rimane segreta
- Si possono *firmare* e *autenticare* i propri messaggi in modo certo e inequivocabile
- Si possono usare per *autenticare* individui o processi in modo certo e inequivocabile

PKC: A scrive a B

- Alice cifra il testo con la chiave *pubblica* di Bob (B_p)
- Bob applica la propria chiave *segreta* B_s e decifra il messaggio di Alice
- Solo Bob possiede la chiave inversa B_s, l'unica che può decifrare il messaggio cifrato da Alice

PKC: A scrive a tutti, firmandosi

- Alice cifra il testo con la propria chiave *segreta* A_s
- Chiunque può decifrare il messaggio di Alice, usando la sua chiave *pubblica* A_p
- Solo Alice può essere l'autrice del messaggio, perché nessun altro conosce la sua chiave segreta A_s

PKC: A scrive a B, firmandosi

- Alice cifra il testo dapprima con la propria chiave *segreta* A_s, poi di nuovo con la chiave *pubblica* di Bob B_p
- Bob decifra il testo dapprima con la propria chiave *segreta* B_s poi di nuovo con la chiave *pubblica* di Alice A_p
- Solo Alice può aver scritto il messaggio, e solo Bob può leggerlo

PKC: A vuole autenticare B

- Alice genera un testo casuale T, lo cifra con la propria chiave *segreta* A_s e lo invia a "Bob"
- Bob decifra il testo con la chiave *pubblica* di Alice A_p poi cifra il risultato con la propria chiave *segreta* B_s e lo restituisce ad Alice
- Alice decifra il testo con la chiave *pubblica* di Bob B_p poi confronta il risultato con T
- Se Alice ha ottenuto T allora il suo interlocutore non può che essere Bob

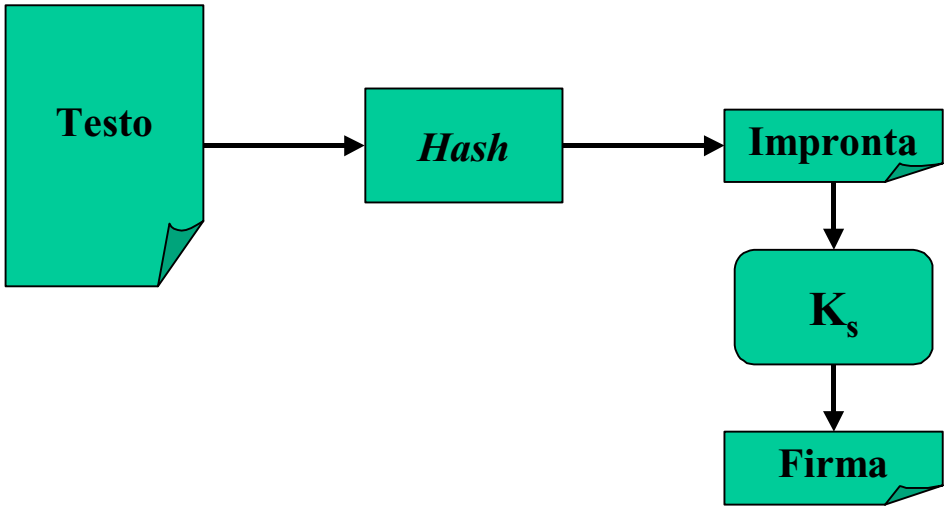
La firma digitale

Concetti di base

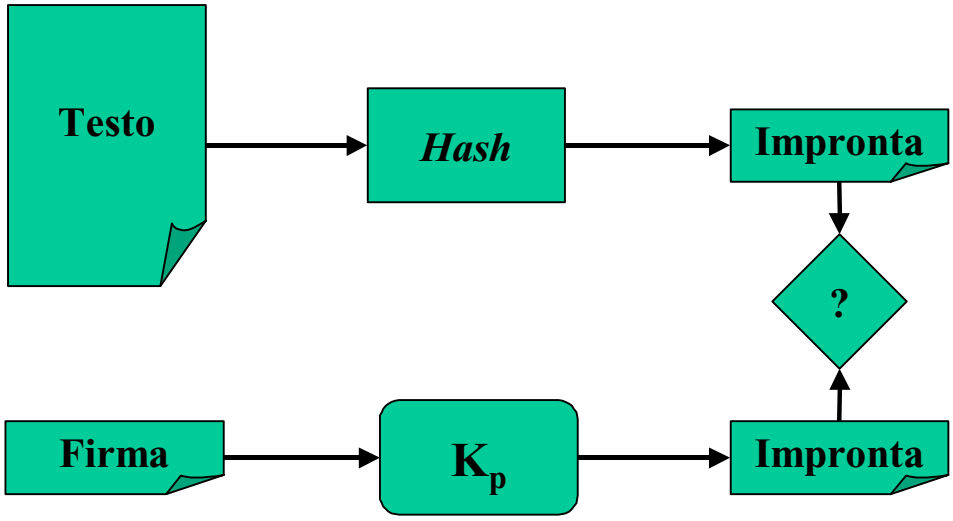
Autenticazione dei documenti

- La “firma digitale”:
  - è verificabile da chiunque
  - non è falsificabile
  - non è ripudiabile
  - rivela modifiche al testo originale
- Prerequisiti:
  - un sistema di cifratura a chiave pubblica
  - una “funzione hash” pubblica
  - un meccanismo di garanzia per le chiavi

Firma digitale: creazione



Firma digitale: verifica



Proprietà della firma digitale

- È il risultato di un calcolo sul documento:
  - Può essere separata dal documento cui si riferisce
  - Non modifica il documento cui si riferisce
- Dipende dal contenuto del documento:
  - Non può essere falsificata, imitata, ripetuta
  - Non può essere apposta “in bianco”
  - Rivela modifiche al documento cui si riferisce
- È verificabile da chiunque
- Non può essere ripudiata

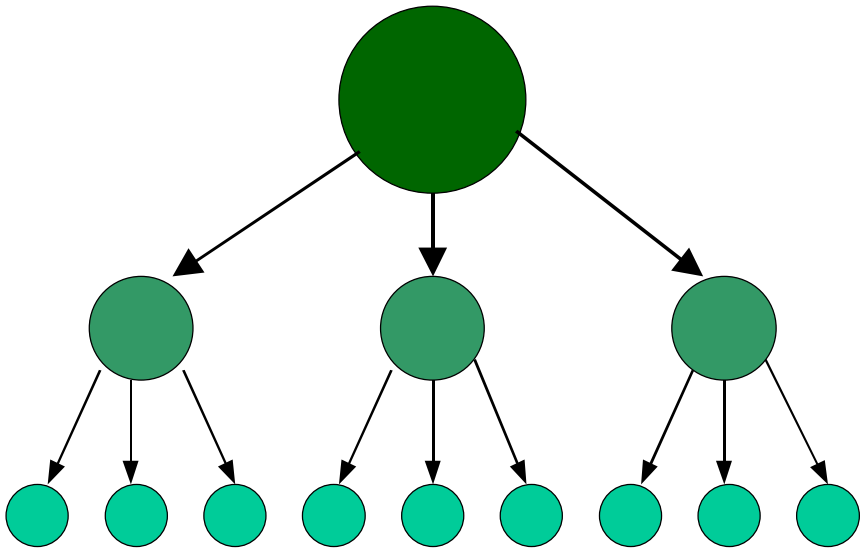
La certificazione

Meccanismi di garanzia sulla validità delle chiavi di firma

Il processo di certificazione

- Perché la firma digitale sia affidabile serve la certezza dell'autenticità delle chiavi pubbliche
- La Autorità di Certificazione (CA):
  - garantisce la validità di una chiave
  - certifica l'identità del relativo titolare
  - gestisce l'elenco delle chiavi pubbliche
- Il Certificato Digitale contiene:
  - la chiave pubblica del titolare "firmata" dalla CA
  - ulteriori dati (scadenza, limitazioni, …)

PKI secondo X.509



I certificati digitali

- “Credenziali elettroniche” che autenticano il titolare di una coppia di chiavi di firma
- Emessi dalle Autorità di Certificazione
- Usati dai browser per:
  - fare transazioni sicure
  - cifrare messaggi di e-mail
  - garantire della provenienza di software
- Danno validità legale ad una firma digitale

Termine dell’intervento

Privacy e firma digitale

WORKSHOP

Connessione in rete: sicurezza informatica e riservatezza

La sicurezza nella Rete Unitaria

M. Terranova

Centro Tecnico per la Rete Unitaria della Pubblica Amministrazione

LA SICUREZZA NELLA RETE UNITARIA

Mario Terranova

Centro Tecnico
per la Rete Unitaria della PA

V. Isonzo, 21b

m.terranova@ct.rupa.it

5ª Conferenza di Statistica

17 novembre 2000

SOMMARIO

- **La Rete Unitaria ed il Centro Tecnico**
- **La sicurezza nella Rete Unitaria**
- **La certificazione delle chiavi per la firma digitale**

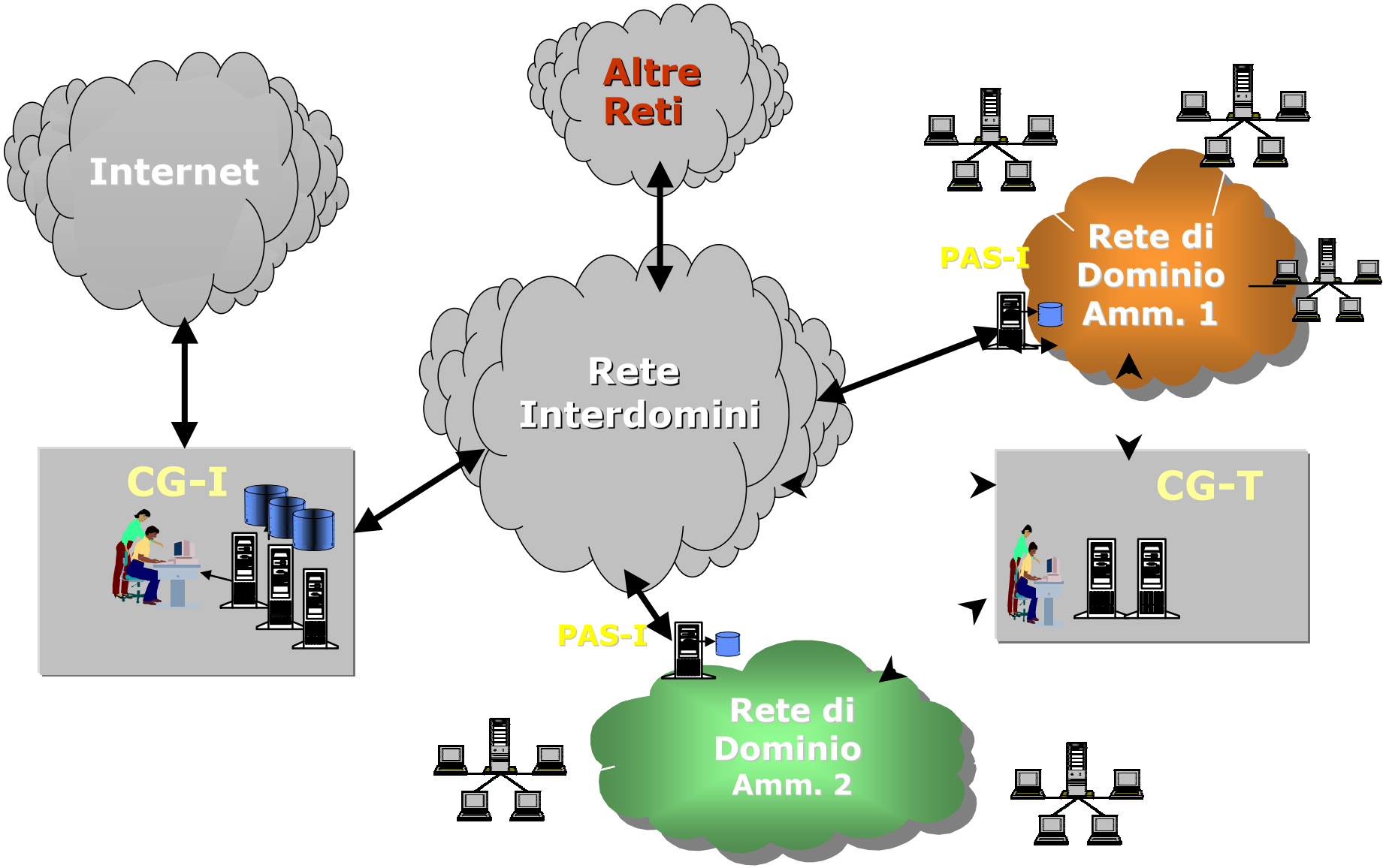
IL CENTRO TECNICO

- **Assiste le Pubbliche Amministrazioni nella stipula, attuazione e gestione dei contratti RUPA.**

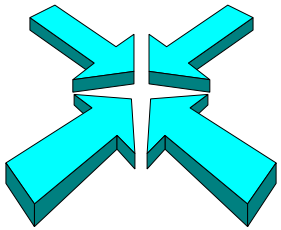
- **Vigila sui livelli di servizio, sulla sicurezza, sulla qualità e l'osservanza delle prestazioni contrattuali.**

- **Svolge le funzioni di certificazione delle chiavi, Notariato e Validazione temporale per gli utenti della rete.**

LA RETE UNITARIA



I SERVIZI DI TRASPORTO

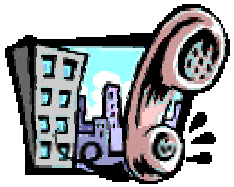


Servizi di connettività e interconnessione Bande di accesso da 64 Kbps fino a 34 Mbps:

- Servizi di connettività IP
- Frame Relay
- X25
- CTN
- IPSec

Servizi di gestione e supporto:

- Controllo dei parametri contrattualizzati
- Call Center (help-desk 2° livello)
- Supporto sistemistico
- Reportistica
- Formazione



I SERVIZI PER L'INTEROPERABILITÀ

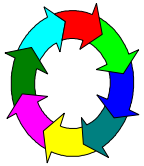
Servizi di interoperabilità trai Domini delle Amministrazioni e con l'esterno:

- Posta elettronica
- Trasferimento file (FTP)
- Terminale virtuale (Telnet)
- Accesso a World WideWeb (WWW)
- Collegamento ad Internet

CG-I

Servizi di gestione e supporto:

- Supporto alla gestione della rete
- Formazione e informazione
- Controllo dei parametri contrattualizzati in aderenza all'accordo quadro



Servizi addizionali per la singola Amministrazione:

- Outsourcing di servizi di pertinenza dell'Amministrazione

SGSA

SOMMARIO

- **La Rete Unitaria ed il Centro Tecnico**
- **La sicurezza nella Rete Unitaria**
- **La certificazione delle chiavi per la firma digitale**

STRUTTURA PER LA SICUREZZA

- **Due livelli distinti: Trasporto ed Interoperabilità**
- **Strutture organizzative per la sicurezza in ciascun livello**
- **Controllo e supervisione esercitata dalla struttura per la sicurezza del Centro Tecnico attraverso:**
  - Revisione dei Piani per la sicurezza
  - Audit
  - Test

SICUREZZA DEL TRASPORTO

- **Trasparenza del CG-T rispetto ai flussi dei dati**
  - il CG-T ha solo funzioni di controllo e gestione della rete di trasporto
  - i dati non attraversano il CG-T
- **Protezioni fisiche e logiche del CG-T**
- **Protezioni fisiche e logiche degli apparati di rete**
- **Servizi di comunicazione sicura (IPSec)**

SICUREZZA DELL'INTEROPERABILITÀ

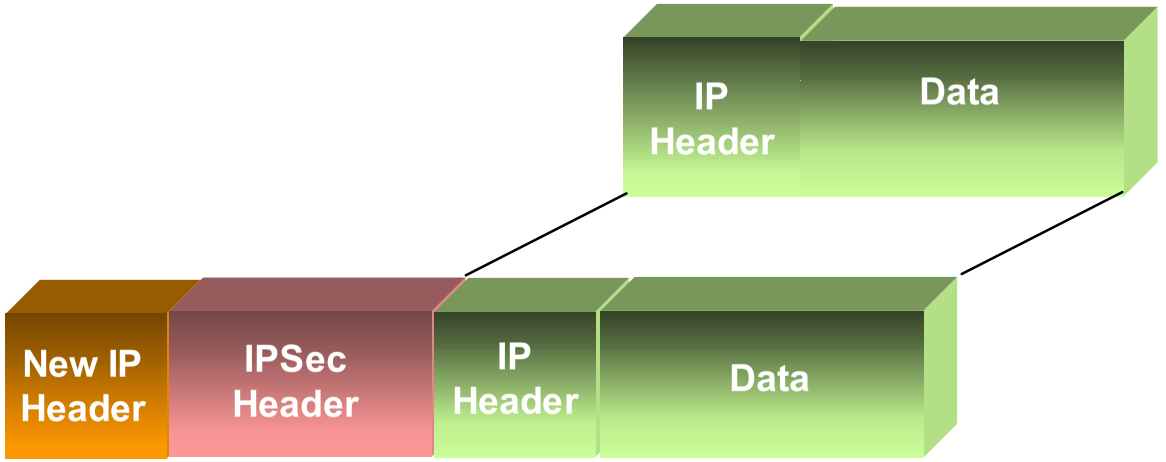
- **Garantire riservatezza ed integrità delle informazioni gestite ed in transito**
- **Impedire a terzi di accedere o modificare dati e risorse di pertinenza delle Amministrazioni**
- **Impedire a personale interno e delle Amministrazioni di accedere o modificare dati e risorse senza averne autorizzazione**
- **Indirizzare al corretto destinatario le informazioni gestite**

IL SERVIZIO IPSec

- **Disponibilità di accessi da 64 Kbit/s fino a 2Mbit/s**
- **Supporto di connessioni ISDN: accesso IP commutato fisso e alta disponibilità**
- **Possibilità di realizzare VPN all'interno di un gruppo chiuso di un dominio**
- **Tunnel multipli**
- **Nested tunnel**

IPSec TUNNEL MODE

- **Protezione delle informazioni**
- **Mascheramento dell'indirizzo IP sorgente**
- **Possibilità di connessione tra LAN con indirizzo non congruente con quello della rete di trasporto**



PROTEZIONE FISICA

- **Controllo accessi con autenticazione multipla:**
  - Smart card individuale
  - PIN
  - Elementi biometrici
- **Sorveglianza del perimetro e dei varchi tramite sistema di videocontrollo con registrazione continua**
- **Allarme perimetrale con sistema combinato motion detector, barriera microonde, infrarosso attivo/passivo, antisfondamento**
- **Sistema di allarme connesso al presidio di sicurezza**

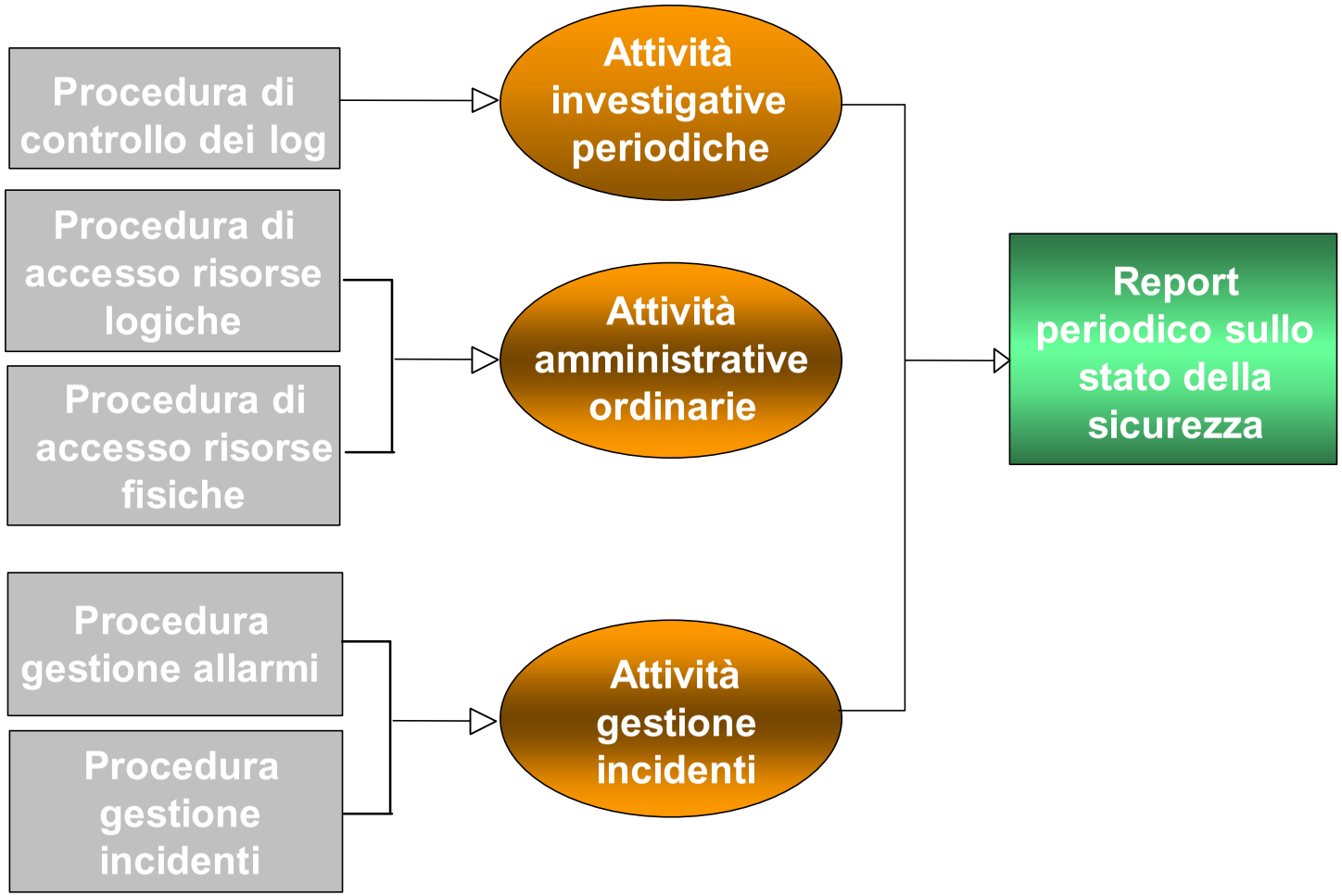
PROTEZIONE LOGICA

- **Profili differenziati di accesso ai sistemi**
- **ACL sugli apparati di rete**
- **Autenticazione forte degli utenti privilegiati con one-time-password**
- **Antivirus**
- **Sistemi di rilevazione delle intrusioni**
- **Protezione dei log**
- **Correlazione eventi**

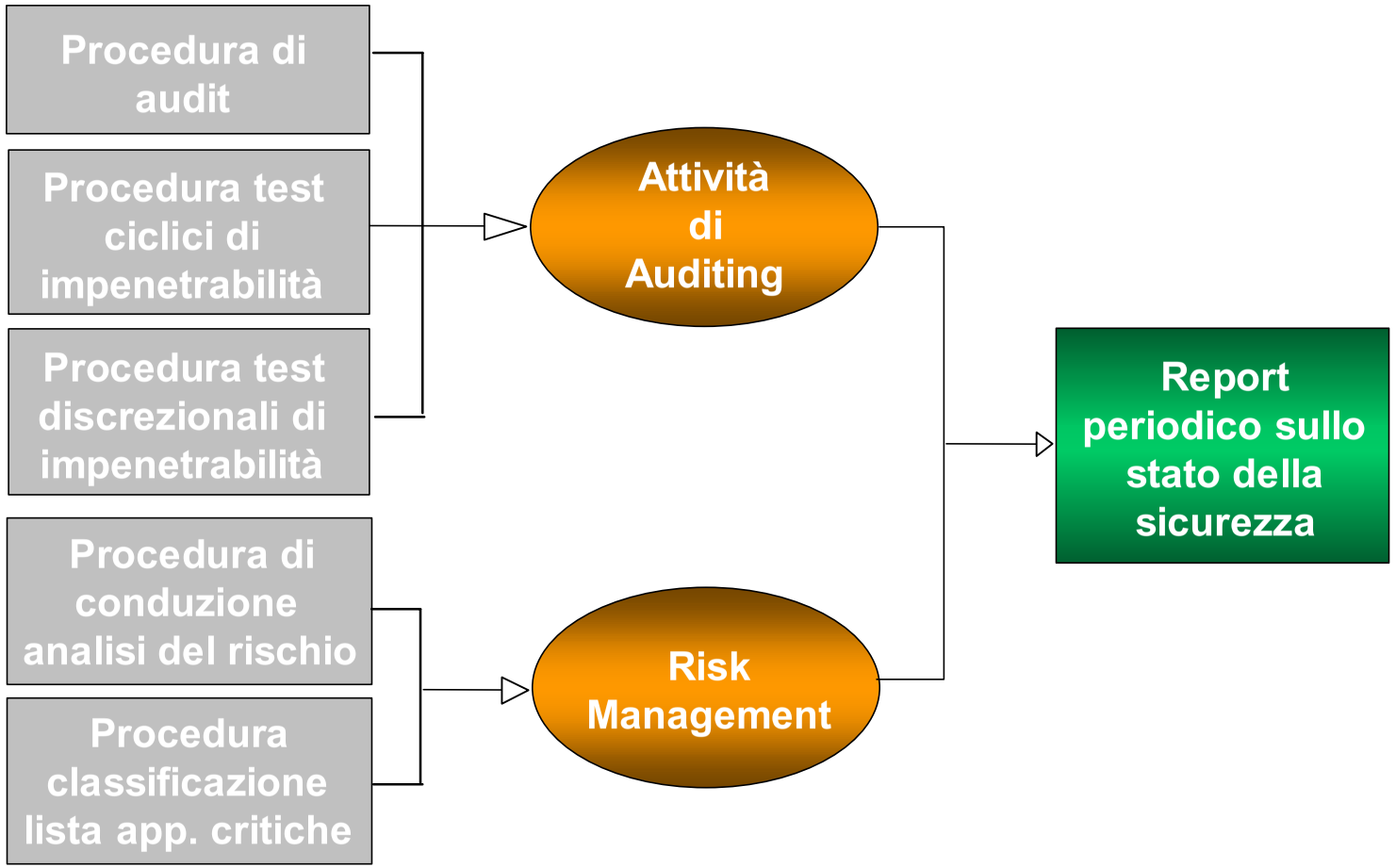
ANALISI DEL RISCHIO

- **Esecuzione dell'analisi del rischio:**
  - In fase di progettazione
  - Su base annua
  - In seguito a gravi incidenti di sicurezza
- **Tre metodi di calcolo:**
  - Riduzione del rischio su parametri BS-7799
  - Riduzione del rischio su parametri tecnologici
  - Efficacia delle barriere su modello topologico

ESERCIZIO DEI DISPOSITIVI DI SICUREZZA



VERIFICHE DI SICUREZZA



TEST DI IMPENETRABILITÀ

- **Esecuzione periodica di test di impenetrabilità**
  - Sul sistema nel suo complesso da:
    - **Internet**
    - **Domini delle amministrazioni**
  - Su singole macchine dalla stessa subnet

- **Impiego di strumenti commerciali e public domain per analisi della sicurezza globale e delle singole applicazioni**

- **Test condotti sia dal Centro Tecnico che da ditte specializzate**

SOMMARIO

- **La Rete Unitaria ed il Centro Tecnico**
- **La sicurezza nella Rete Unitaria**
- **La certificazione delle chiavi per la firma digitale**

IL QUADRO NORMATIVO DI RIFERIMENTO

- **DPR 23 dicembre 1997, n. 522, articolo 2, comma 2, lettera d).**

  Affidamento al Centro Tecnico delle attività di certificazione per gli utenti della Rete Unitaria

- **DPCM 8 febbraio 1999, art. 16, comma 5.**

  Prevede l'iscrizione del Centro Tecnico nell'Elenco Pubblico dei certificatori

- **DPCM 8 febbraio 1999, art. 62.**

  Permette alle pubbliche amministrazioni di svolgere l'attività di certificatore utilizzandoi servizi di un certificatore

IL CONTRATTO DI FACILITY MANAGEMENT

- **Contratto di quattro anni con POSTECOM**
- **Infrastruttura di certificazione**
  - 30.000 dispositivi di firma
  - 60.000 coppie di certificati
  - 60.000 smart card
  - 10.000 marche temporali
- **Inizio delle attività entro la prima decade di dicembre**
- **Le Amministrazioni devono adeguare i procedimenti amministrativi al fine di consentire l'uso efficiente della Firma digitale**

CONCLUSIONI

- **La sicurezza è il risultato di:**
  - misure tecnologiche, procedurali ed organizzative
  - processo continuo di analisi e verifica

WORKSHOP

Connessione in rete: sicurezza informatica e riservatezza

Protezione della riservatezza in basi di dati

E. Bertino

Protezione della Riservatezza in Basi di Dati

Elisa Bertino

Rilascio di Informazioni

- Il rilascio dei dati provoca divulgazione quando dati che dovrebbero essere trattati come confidenziali sono resi noti

- La divulgazione può:

  - essere basata solamente sui dati rilasciati

  - risultare dalla combinazione di dati rilasciati con informazione pubblicamente disponibile

Metodi di rilascio - tabelle

- Ogni entrata in una tabella rappresenta il valore aggregato di una quantità su tutte le unità di analisi che appartengono ad un'unica cella

- Due categorie di tabelle:

  tabelle di frequenza (count)

  tabelle di grandezze (magnitude data)

Tabella di frequenza - esempio

Esempio - numero di beneficiari di un certo benefit per importo e per contea

Importo

Contea	$0-19	$20-39	$40-59	$60-79	$80-99	$100+	Totale
A	2	4	18	20	7	1	52
B	-	-	7	9	-	-	16
C	-	6	30	15	4	-	55
D	-	-	2	-	-	-	2

Metodi di rilascio - microdati

Un file di microdati è composto da record individuali ognuno contenente valori di variabili per una singola persone, entità commerciale, o altra unità

File di microdati - esempio

Esempio - tutti i record della contea Alfa relativi a bambini delinquenti

N	Bambino	Contea	Educ. CF	Stipendio CF	Razza CF
1	John	Alfa	Molto alto	201	nera
2	Jim	Alfa	Alto	103	bianca
3	Sue	Alfa	Alto	77	nera
4	Pete	Alfa	Alto	61	bianca
5	Ramesh	Alfa	Medio	72	bianca
6	Dante	Alfa	Basso	103	bianca
7	Virgil	Alfa	Basso	91	nera
8	Wanda	Alfa	Basso	84	bianca
9	Stan	Alfa	Basso	75	bianca
10	Irmi	Alfa	Basso	62	nera
11	Renee	Alfa	Basso	58	bianca
12	Virginia	Alfa	Basso	56	nera
13	Mary	Alfa	Basso	54	nera
14	Kim	Alfa	Basso	52	bianca
15	Tom	Alfa	Basso	55	nera
16	Ken	Alfa	Basso	48	bianca
17	Mike	Alfa	Basso	48	bianca
18	Joe	Alfa	Basso	41	nera
19	Jeff	Alfa	Basso	44	nera
20	Nancy	Alfa	Basso	37	bianca

Divulgazione di informazione

- Divulgazione è l'attribuzione impropria di informazione ad un oggetto, sia esso individuo o organizzazione

- Tipi di divulgazione:

  divulgazione di identità

  divulgazione di attributo

  divulgazione inferenziale

Dati ristretti ed accesso ristretto

La confidenzialità può essere protetta

- restringendo la quantità di informazione nelle tabelle o file di microdati - dati ristretti
- imponendo condizioni sull'accesso ai dati prodotti - accessi ristretti
- combinazione dei due metodi precedenti

Protezione di tabelle di frequenza

- I dati raccolti da molti survey sulla popolazione sono pubblicati in tabelle che presentano somme (numero di persone per categoria) o frequenze (frazioni o percentuali di persone per categorie)

- Tecniche di protezione includono

  - regole di soglia

  - regole speciali

Regole di soglia

- Una cella di una tabella è definita sensitiva se il numero dei rispondenti è minore di un certo numero specificato (ad esempio 5, oppure 3)
- Le celle sensitive non possono essere rilasciate
- Per soddisfare tale regola, si possono usare varie tecniche

  - soppressione di celle
  - ristrutturazione di tabelle e combinazione di categorie
  - arrotondamento casuale
  - inserimento confidenziale

Tabella con divulgazione

Esempio - numero di bambini delinquenti per contea e livello di educazione del capo famiglia

celle con meno di 5 rispondenti sono considerate *sensitive*

Livello di educazione

Contea	**Basso**	**Medio**	**Alto**	**Molto Alto**	**Totale**
Alfa	15	**1**	**3**	**1**	20
Beta	20	10	10	15	55
Gamma	**3**	10	10	**2**	25
Delta	12	14	7	**2**	35
Totale	50	35	30	20	135

Soppressione di celle

- Una delle tecniche maggiormente diffuse è quella della soppressione di celle
- L'uso della soppressione primaria non è tuttavia spesso sufficiente a garantire protezione
- Almeno un'altra cella deve essere soppressa - soppressione complementare - per ogni riga e colonna nella quale è stata soppressa una cella sensitiva
- Anche in presenza di soppressione complementare è difficile garantire che il risultato fornisca protezione adeguata

Tabella con divulgazione non protetta da soppressione

Esempio - numero di bambini delinquenti per contea e livello di educazione del capo famiglia

Livello di educazione

Contea	**Basso**	**Medio**	**Alto**	**Molto Alto**	**Totale**
Alfa	15	**D1**	**D2**	**D3**	20
Beta	20	D4	D5	15	55
Gamma	**D6**	10	10	**D7**	25
Delta	D8	14	7	**D9**	35
Totale	50	35	30	20	135

Sembra garantire protezione ma…….

(15 + D1 + D2 + D3) + (20 + D4 + D5 + 15) - (D1 + D4 + 10 + 14) -
(D2 + D5 + 10 + 7) = 20 + 55 - 35 - 30

===> D3 =1

Soppressione di celle

- La scelta delle celle da eliminare con soppressione complementare è più complicata di quanto appare
- Tecniche di programmazione lineare sono utilizzate per la scelta automatica di celle da eliminare con soppressione complementare
- Tecniche di audit possono essere utilizzate per valutare la bontà della soppressione per vedere se fornisce protezione adeguata

Tabella con divulgazione protetta da soppressione

Esempio - numero di bambini delinquenti per contea e livello di educazione del capo famiglia

Livello di educazione

Contea	**Basso**	**Medio**	**Alto**	**Molto Alto**	**Totale**
Alfa	15	**D**	**D**	**D**	20
Beta	20	10	10	15	55
Gamma	**D**	D	10	**D**	25
Delta	D	14	D	**D**	35
Totale	50	35	30	20	135

La tabella garantisce protezione, ma……delle 16 celle componenti la tabella solo 7 sono pubblicate, mentre 9 sono soppresse

Regole speciali

Impongono restrizioni sul livello di dettaglio che può essere riportato in una tabella

Esempio - Social Sec. Administration proibisce tabelle nelle quali i dettagli di una cella sono uguali al totale parziale o permetterebbero all'utente di determinare l'età di un individuo all'interno di un intervallo di 5 anni, stipendi all'interno di un intervallo di $1000 o benefits all'interno di un intervallo di $50

Regole speciali

Esempio - numero di beneficiari per importo e contea

	Importo						
Contea	**$0-19**	**$20-39**	**$40-59**	**$60-79**	**$80-99**	**$100+**	**Totale**
A	2	4	18	20	7	1	52
B	-	-	7	9	-	-	16
C	-	6	30	15	4	-	55
D	-	-	2	-	-	-	2

Non può essere pubblicata. I dati per B e D violano le regole
- per D esiste una sola cella non vuota, la cella dettagliata è uguale al totale
- per B ci sono due celle non vuote ma l'intervallo dei possibili benefit è tra $40 e $79. Tale intervallo è più piccolo dei $50 consentiti

Regole speciali

- per evitare divulgazione non autorizzata i dati delle contee possono essere combinati
- la combinazione di categoria ("rolling-up" categories) può essere effettuata per righe e per colonne

	Importo						
Contea	**$0-19**	**$20-39**	**$40-59**	**$60-79**	**$80-99**	**$100+**	**Totale**
A&B	2	4	25	29	7	1	68
C&D	-	6	32	15	4	-	57

Microdati

- un metodo alternativo alle tabelle per pubblicare risultati raccolti da censimenti o survey è quello di rilasciare file di microdati di uso pubblico
- sempre più spesso oggi i dati sono rilasciati in forma di microdati
- i microdati risultano più convenienti per i riceventi dell'informazione che possono calcolare statistiche in base ai loro bisogni

Microdati

- È noto o che la protezione dei microdati è più complessa della protezione di tabelle
- una delle principali difficoltà nella protezione dei microdati è la possibilità di combinare l'informazione rilasciata con altra informazione disponibile
- non esistono misure largamente accettate per ridurre il rischio di divulgazione per i microdati, quindi non esistono "standard" che potrebbero essere applicati per valutare se una protezione è adeguata

Tecniche di protezione per microdati

Per ridurre la possibilità di divulgazione, la maggior parte dei file di microdati rilasciati:

- include dati solo di un campione della popolazione
- non include identificatori espliciti
- limita i dettagli geografici
- limita il numero di variabili del file

Tecniche di protezione per microdati

Metodi addizionali includono:

- codifica al massimo e minimo
- aggiunta o moltiplicazione per numeri casuali (noise)
- scambio (swapping)
- registrazione per intervalli o arrotondamenti
- aggregazione per piccoli gruppi di rispondenti e sostituzione dei valori individuali con la media (blurring)

Basi di dati statistiche

- Un DB statistico mantiene i dati on-line in un DBMS (generalmente relazionale) che effettua controlli di inferenza dinamicamente
- le relazioni in un DB statistico sono simili ai file di microdati ma mantengono specifiche informazioni sensitive, quali stipendio, che possono essere rilasciate solo in forma statistica
- molti DB statistici sono usati solo per scopi statistici; altri sono utilizzati per scopi più generali, permettendo sia accesso statistico che non statistico

Query statistiche

- Le query statistiche richiedono l'applicazione di funzioni statistiche (sum, count, median, avg) applicate a sottoinsiemi della base di dati
- la dimensione del query set è il numero di record ai quali la funzione è applicata
- l'insieme dei record ai quali applicare la funzione è specificato attraverso una funzione caratteristica, che è una formula logica sui valori degli attributi
- esempio: Funzione caratteristica

  C1 = (sex=Male) AND ((Major=CS) OR (Major = EE))

  |C1| = 6

Basi di dati statistiche

- La risposta ad una o più query statistiche può divulgare informazione riservata
- l'inferenza che un utente può effettuare è spesso basata su conoscenza esterna
- per proteggere la riservatezza, il DBMS può restituire la risposta precisa, una risposta approssimata, o non restituire alcuna risposta
- uno degli ostacoli principali nel proteggere le basi di dati da divulgazione è dato dall'inferenza che potrebbe essere effettuata mediante la sovrapposizione di più richieste

Statistiche sensitive

- Una statistica è sensitiva se divulga informazione confidenziale circa entità rappresentate nel DB

- una statistica su un query set di dimensione 1 è sempre sensitiva

- una statistica calcolata su un insieme di due elementi può essere considerata sensitiva, perché un utente con conoscenza addizionale circa uno dei due valori potrebbe inferire l'altro

Statistiche sensitive

- Un'altra possibile politica è basata sulla la regola (n,k) che per *n*-rispondenti *k%*-dominazione, definisce sensitiva una statistica se n o meno dei valori contribuisce al *k%* del totale
- **esempio**: statistica che dà la somma delle entrate di IBM e di qualche piccolo produttore potrebbe essere sensitiva alla regola "1-rispondente 99%-dominazione"
- Il suo rilascio potrebbe fornire informazioni sulle entrate dell'IBM (cifre più significative)

Misure di protezione

- Controllo sulla dimensione della query - in generale query su set piccoli sono proibite
- Se il linguaggio di query permette l'utilizzo di negazioni, allora anche query set molto grandi devono essere proibiti, altrimenti gli utenti potrebbero inferire informazioni usando NOT C
- In generale una query statistica q(C) è permessa solo se

  $$n \leq |C| \leq N-n$$

  dove n è un parametro (intero positivo) fissato

Conclusioni

- Le specifiche misure di protezione dei dati dipendono dalla specifica tecnica usata per rilasciare i dati
- Sono state definite numerosissime tecniche
- Problemi aperti: protezione dei dati a fronte di analisi tramite tecniche di data mining

WORKSHOP

Connessione in rete: sicurezza informatica e riservatezza

Sicurezza informatica e produzione statistica

A. Guarino

SICUREZZA INFORMATICA
E PRODUZIONE STATISTICA

Andrea Guarino

Roma, 17 novembre 2000

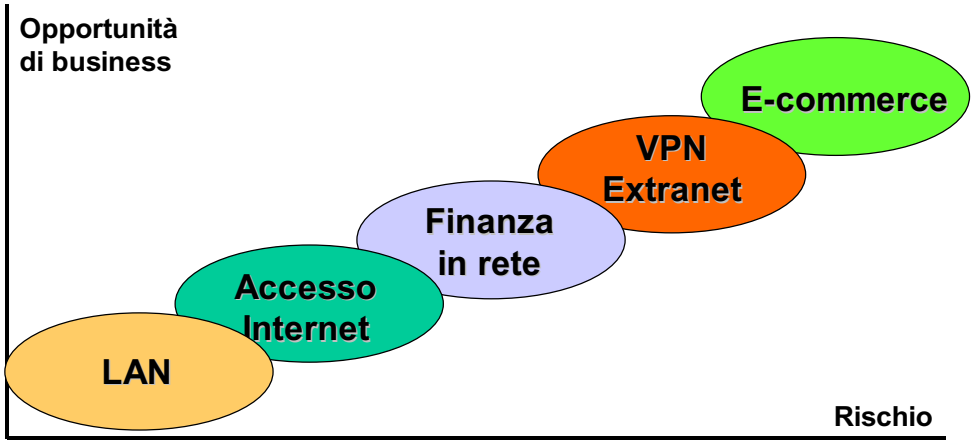
ISTAT

Lo scenario

- Estensione delle opportunità
- Aumento del rischio
- Evoluzione delle tecniche di sicurezza
- Evoluzione delle tecniche di produzione statistica

Estensione delle opportunità

Interazione con gli utenti
sempre più spinta



Evoluzione della sicurezza

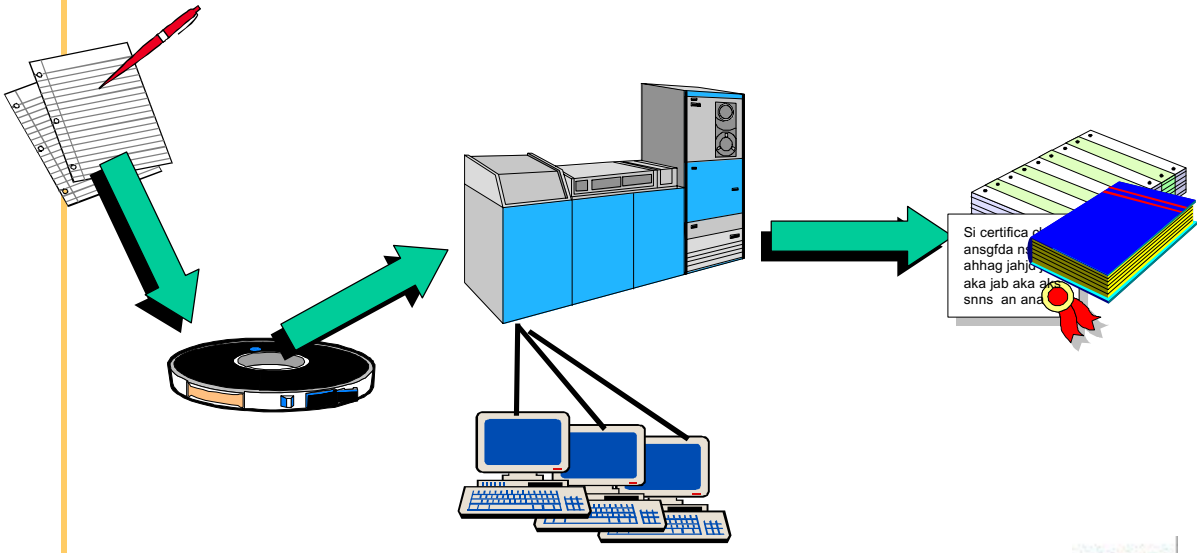
Da una politica di divieti
ad una politica di permessi

Gli utenti interni accedono alle risorse esterne
E' vietato l'accesso dall'esterno

Permettere agli utenti esterni
di accedere a servizi e risorse

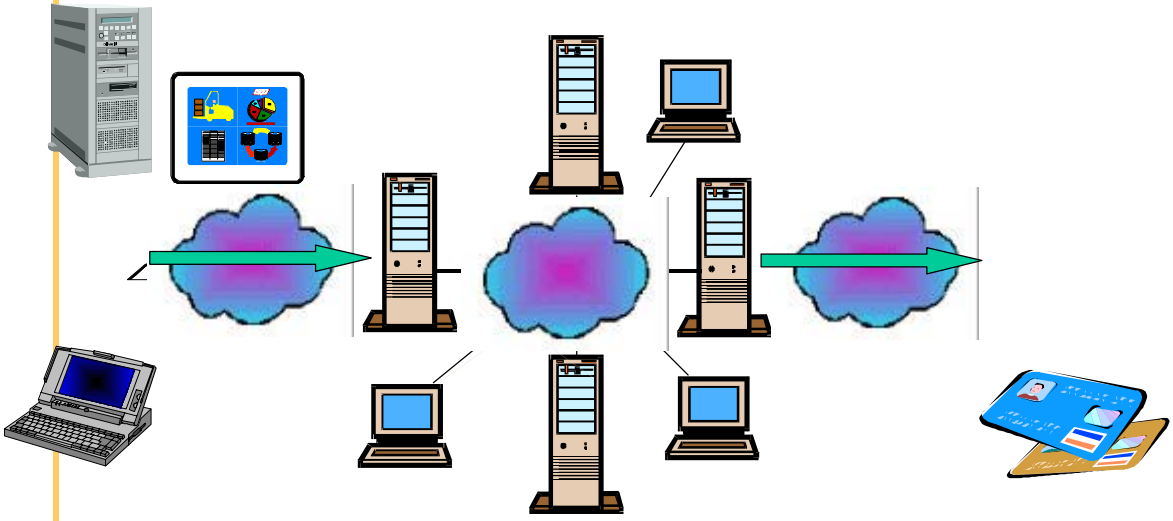
Produzione statistica

Dalla carta ...



Produzione statistica

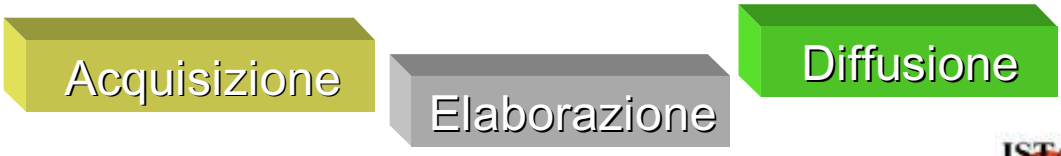
... all'informazione digitale



Produzione statistica

- Da "*host-centric*" a "*network-centric*"
- Adozione di nuovi strumenti e nuove tecniche
- Innalzamento del livello di sicurezza

Nuove modalità di produzione



Acquisizione

Dalla carta all'acquisizione diretta dell'informazione digitale

- Form
- E-mail
- Trasferimento file
- Questionari

- **Autenticazione**
- **Integrità**
- **Riservatezza**

Acquisizione

Dalla carta all'acquisizione diretta dell'informazione digitale

- Form
- E-mail
- Trasferimento file
- Questionari

- **PKI**
- **SSL**
- **IPsec**

Elaborazione

Dall'informazione su un elaboratore all'informazione in rete

- DB condivisi in rete
- Elaborazione a distanza

- **Autenticazione**
- **Riservatezza**

Elaborazione

Dall'informazione su un elaboratore all'informazione in rete

- DB condivisi in rete
- Elaborazione a distanza

- **PKI**
- **IPsec**

Diffusione

L'informazione digitale diventa informazione primaria

- WEB
- Applet
- Documenti legali
- E-book

- **Autenticazione**
- **Integrità**
- **Validità legale**

Diffusione

L'informazione digitale diventa informazione primaria

- WEB
- Applet
- Documenti legali
- E-book

- **PKI**
- **SSL**
- **Firma digitale**

Quali implicazioni ?

- Una seria organizzazione interna
- Introduzione di nuove figure professionali
- Introduzione di una nuova cultura : "La sicurezza come bene primario irrinunciabile"
- Definizione di scelte di base: Tutto in casa, In outsourcing, In housing ?
- Definizione di una politica nei confronti dei partner e degli utenti

A quando la cross-certification ?

Quali benefici ?

- Maggiore tempestività
- Maggiore accuratezza
- Snellimento delle procedure
- Estensione dell'utenza

Miglioramento del servizio

Quali costi ?

"Una tassa da pagare in termini economici, organizzativi e prestazionali !"

Da centro di costo a centro di profitto

Un investimento irrinunciabile per cogliere nuove opportunità

QUINTA
CONFERENZA
NAZIONALE
DI STATISTICA

SICUREZZA INFORMATICA
E PRODUZIONE STATISTICA

Ing. Andrea Guarino

Roma , 17 novembre 2000

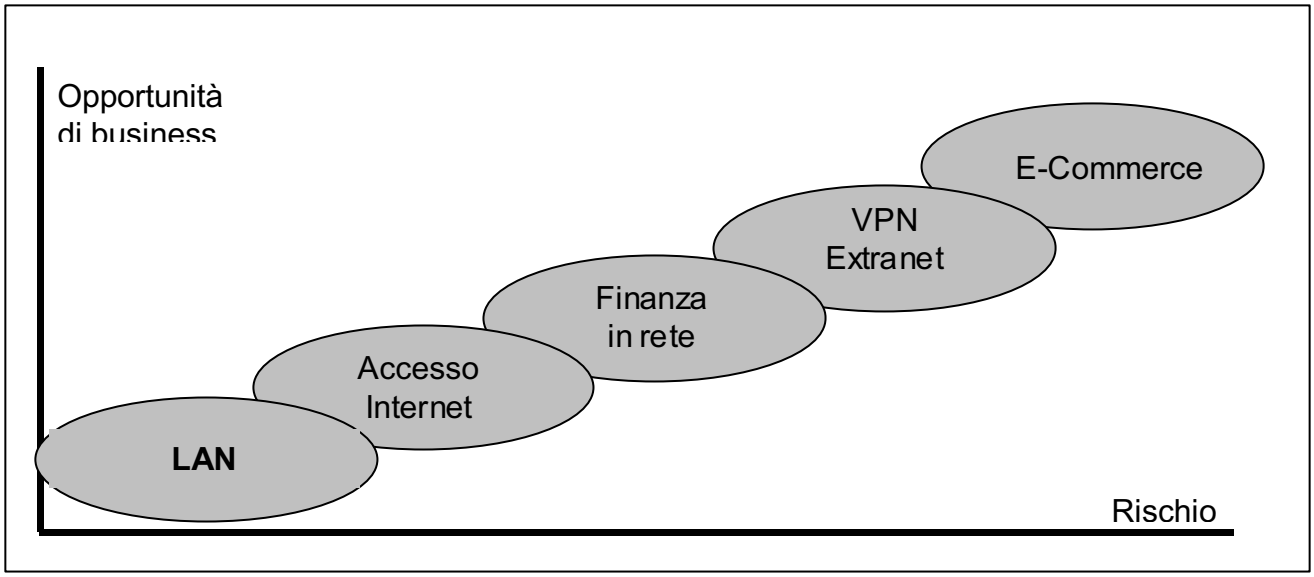
ISTAT

Lo scenario del mondo dell' Information Technology in generale, e della produzione statistica in particolare, si è rapidamente evoluto sotto l' impulso di fattori di spinta dalla portata dirompente che hanno permesso di ripensare al rapporto tra utente e sistemi informatici in modo completamente diverso.

L' affermarsi dell'informatica individuale con una diffusione sempre più spinta di sistemi di elaborazione personali dalle caratteristiche in termini di prestazioni impensabili fino a qualche anno fa e soprattutto il prorompente sviluppo delle reti di comunicazioni sia in termini di qualità che di diffusione, hanno permesso di accorciare la distanza tra di essi, ed ancora di più quella tra cittadini e Pubblica Amministrazione. Viene definito infatti un nuovo modello di interazione tra di essi ed un nuovo modo di usufruire dei servizi che la Pubblica Amministrazione, finalmente unica, riesce ad offrire.

Si è assistito così alla nascita di servizi di qualità via via crescente offerti da quelle organizzazioni che hanno saputo cogliere le nuove opportunità di business che si sono venute a creare. Ma non si può pensare a tali opportunità senza dover affrontare il problema della sicurezza e soprattutto senza dover valutare il rischio cui ciascuna organizzazione è esposta nel momento in cui decide di offrire nuovi servizi ad un' utenza sempre più vasta.

Volendo ripercorrere le veloci tappe bruciate in questi ultimi anni si può notare infatti come l' offerta di servizi informatici abbia coperto utenze via via più diffusa a fronte di un crescita esponenziale del rischio di danneggiamento di risorse ed immagine.

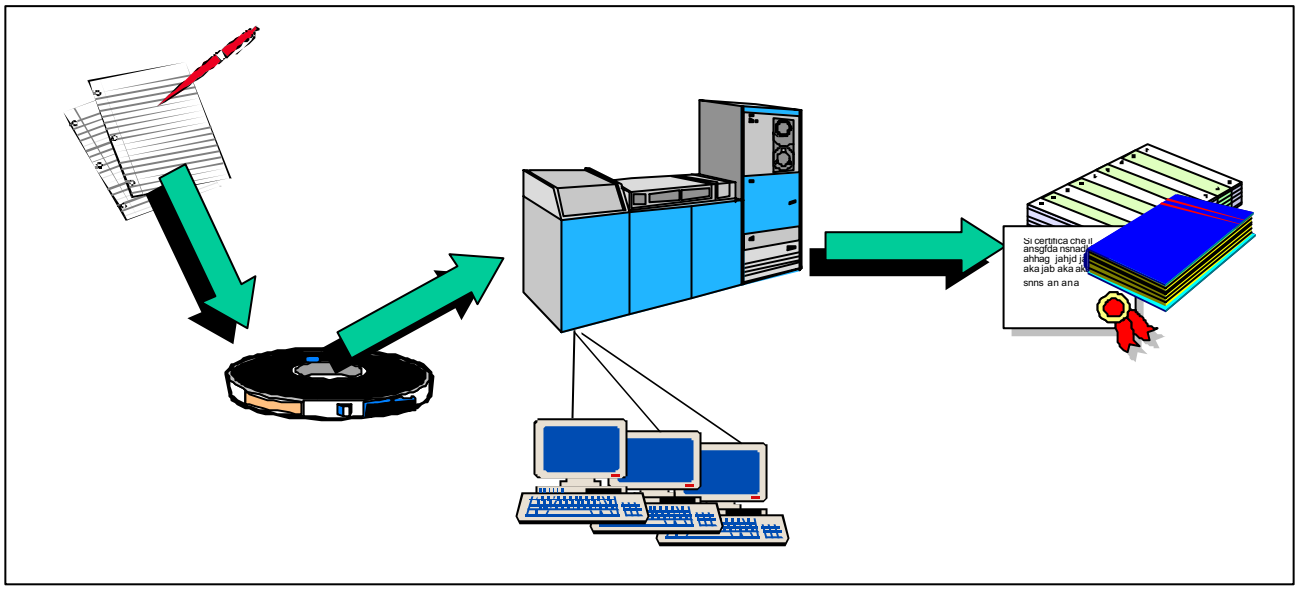


Dal rischio quasi nullo corrispondente all' offerta di servizi su reti chiuse quale poteva essere una rete locale, in cui era lo stesso utente a dover recarsi presso il fornitore del servizio, che

rappresentava al tempo stesso l' unico elemento abilitato all' accesso alle risorse ed alle informazioni interne, si è successivamente passati alla prima fornitura di servizi su Internet.
Tali servizi hanno costituito il primo passo verso una diffusione più spinta dell' informazione, anche se essa era un' informazione ancora statica e spesso unicamente di facciata, ma al tempo stesso hanno rappresentato il primo caso di apertura delle reti e di esposizione di risorse su una rete potenzialmente aperta a tutti. E' nato così il problema della protezione delle risorse interne da una parte e dei contenuti esposti su Internet dall' altra; problemi questi che hanno assunto un' importanza sempre crescente man mano che le informazioni pubblicate hanno assunto maggior importanza e gli utilizzatori di Internet sono esponenzialmente aumentati, diventando al tempo stesso sempre più scaltri.
I passi successivi hanno visto il mondo della finanza, prima di ogni altro, accedere alle potenzialità offerte dalla rete offrendo ai propri clienti i primi servizi finanziari in linea a fronte di investimenti nel campo della sicurezza tutt' altro che indifferenti. Ed ancora dopo, a fronte dell' affermarsi di tecniche di sicurezza sempre più sofisticate, alcune organizzazioni hanno compreso gli enormi vantaggi ed opportunità che si venivano a creare a fronte dell' utilizzo della rete Internet come mezzo di comunicazione aziendale e della possibile intima cooperazione con altre realtà: sono nate così le prime Reti Private Virtuali (V.P.N.) su Internet e la realizzazione delle Extranet.
L' ultimo scalino, costituito dal diffondersi attuale del Commercio Elettronico, vede da un lato l' estensione massima dell' utenza (potenzialmente chiunque accede alla rete Internet è un possibile cliente) e dall' altro il raggiungimento di un livello di rischio elevatissimo (vengono esposti i servizi portanti delle singole organizzazioni).
Con il succedersi di queste tappe si è passati quindi da un servizio offerto a pochi che dovevano obbligatoriamente recarsi presso il fornitore ad un servizio offerto a tutti direttamente sul proprio tavolo di lavoro, ma contemporaneamente da un rischio di intrusione minimo ad uno elevatissimo.
Le tecniche di sicurezza hanno quindi dovuto adeguarsi alle necessità crescenti dell' utenza e dei fornitori di servizi, fornendo strumenti sempre più sofisticati che hanno contemporaneamente permesso di pensare a nuove applicazioni.
La sicurezza inizialmente era basata completamente sulla protezione delle risorse interne alle organizzazioni, in cui mentre veniva concesso agli utenti interni di accedere a servizi esterni sulla rete Internet, veniva allo stesso tempo vietata qualsiasi attività da parte di utenti Internet sulle macchine interne.
Ma tale approccio è risultato limitativo e frenante per lo sviluppo dei servizi. Pertanto nel corso del tempo si è passati all' utilizzo di tecniche che pur proteggendo i sistemi interni hanno permesso

l' accesso da parte di utenti Internet alle informazioni ed alle applicazioni proprie delle singole organizzazioni.

Anche la produzione statistica ha subito gli stessi influssi da parte dei medesimi fattori di spinta. In essa, infatti, inizialmente le fasi di acquisizione e di diffusione erano totalmente basate sulla carta.



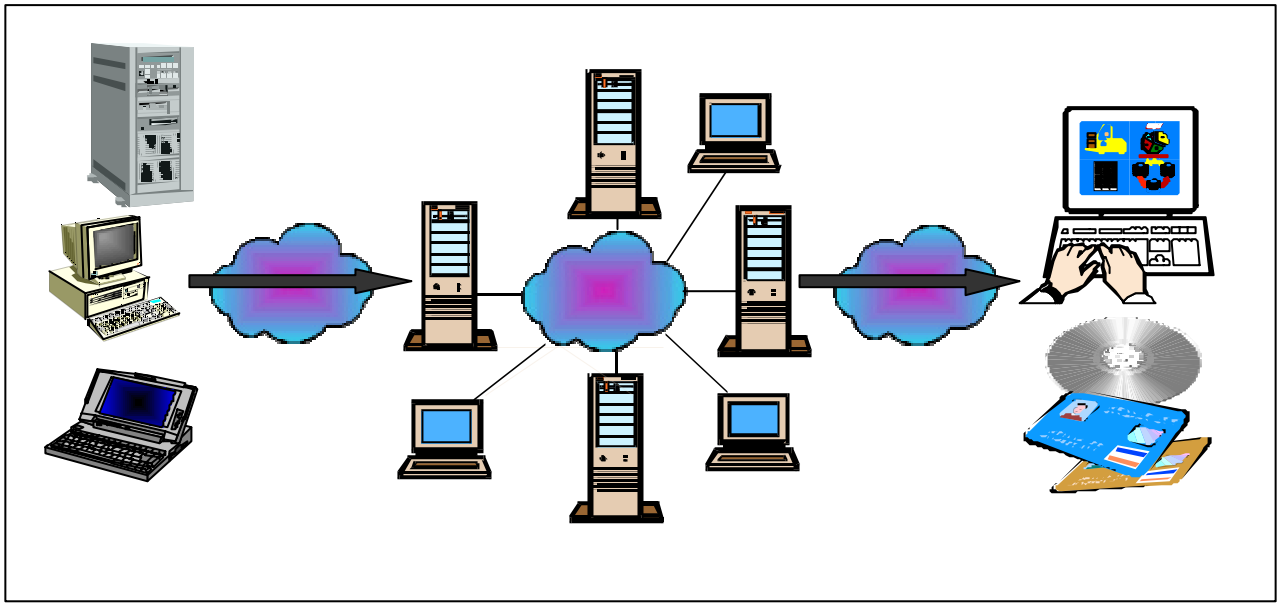
La compilazione di questionari cartacei ha rappresentato per lungo tempo la prima fase del processo di produzione statistica; questi venivano quindi passati al processo di "perforazione" attraverso il quale l' informazione cartacea veniva trasformata in informazione digitale. Un unico elaboratore, cui erano collegati terminali attraverso il quale il processo veniva impostato e controllato, macinava le informazioni elementari per produrre finalmente gli oggetti per la diffusione: tabulati, pubblicazioni, certificati.

Il processo di produzione quindi vedeva una transizione carta –> digitale –> carta che ne rallentava l' esecuzione, introducendo al tempo stesso possibili fonti di errori nella fase di perforazione ed aumentandone i relativi costi.

I primi miglioramenti al processo si sono avuti con l' introduzione delle tecniche di lettura ottica dei questionari e con l' acquisizione di dati direttamente su supporti magnetici. Contemporaneamente ai prodotti cartacei di diffusione si sono affiancati i primi prodotti di informazione digitale attraverso la diffusione in allegato di floppy-disk o CD o attraverso la costituzione delle B.B.S. (Bolletin Board System).

Lo scenario che ci si presenta oggi è completamente diverso: l' informazione digitale diventa la fonte primaria di acquisizione e diffusione.

Essa viene infatti acquisita direttamente alla fonte attraverso un processo di trasferimento di dati tra sistemi connessi in rete, o direttamente introdotta nel sistema "target" da utenti o rilevatori muniti di palm-computer.



L' informazione così acquisita è direttamente disponibile in rete in una struttura di dati organizzata per i settori produttivi.

La stessa architettura di sistema evolve dal modello precedente a quello distribuito dove il carico di lavoro viene suddiviso tra i sistemi di competenza specifica dei singoli settori, con indubbi vantaggi in termini di potenza elaborativa e di snellimento del processo.

Sono infine gli stessi settori produttivi che possono curare la diffusione dell' informazione da essi stessi prodotta direttamente in rete, concedendo l' accesso ad essa ad utenti selezionati. Quest' ultimo passo ha una portata enorme in termini di rinnovamento. E' l' utente stesso, il ricercatore esperto, l' azienda, il professionista o il privato cittadino, che diventa l' attore principale del processo estraendo l' informazione strutturata in base alle proprie esigenze e nel momento in cui più gli interessa.

In questa evoluzione del processo produttivo dell' informazione statistica è la rete ad essere passata al ruolo di elemento principale del sistema (modello *"network centric"*) sostituendo il vecchio modello in cui invece l'elemento principale era costituito dall' elaboratore centrale (modello *"host centric"*).

Evoluzione che è stata resa possibile dal grande sviluppo delle reti di telecomunicazione sia in termini di diffusione che di qualità e dall' affermarsi delle nuove tecniche e dei nuovi strumenti di

sicurezza che hanno permesso da un lato l' apertura a nuovi servizi e dall'altro il complessivo aumento del livello di sicurezza dei sistemi.

Il processo di produzione statistica è dunque profondamente cambiato nelle sue fasi principali di acquisizione, elaborazione e diffusione.

Nella fase di acquisizione la rilevazione di dati su questionari cartacei è stata sostituita da nuove tecniche che, se pur a fronte di un aumento del rischio, comportano un drastico miglioramento del servizio di raccolta sia in termini di tempestività che di accuratezza.

I principali nuovi metodi di acquisizione possono essere raggruppati nelle seguenti categorie:

- FORM: E' la tecnica attraverso la quale vengono resi disponibili dei questionari elettronici direttamente su un sito WEB dedicato alla raccolta dei dati. In questo modo i dati da rilevare, controllati nella fase formale sulla stazione client, arrivano direttamente in una forma strutturata come previsto dal settore di produzione. I principali problemi legati a questo tipo di acquisizione sono quelli relativi all'autenticazione del mittente e del destinatario (E' proprio quel particolare rispondente a spedire i dati ed è proprio il server dell'Istituto a riceverli ?), dell'integrità (I dati giungono non alterati al server ?) e quello della riservatezza (I dati personali sono visibili ad altri ?).
- E-mail: Un'altra modalità di acquisizione a basso costo è quella che utilizza la posta elettronica per la spedizione di dati ed informazioni. In questo caso è necessaria una preventiva fornitura agli utenti di software certificato per l'immissione ed il controllo dei dati. La spedizione degli archivi di dati rilevati come allegati di messaggi di posta elettronica permette di raccogliere a costi molto contenuti enormi quantità di dati, permettendo al tempo stesso la compilazione fuori linea dei questionari. Anche in questo caso i problemi principali che si presentano riguardano la mutua autenticazione, l'integrità e la riservatezza. Ad essi però si aggiunge l'affidabilità del mezzo (E' certo che il destinatario ha ricevuto il messaggio?).
- Trasferimento file: E' una tecnica simile a quella precedentemente descritta ma con la quale il trasferimento delle informazioni tra utenti e Istituto avviene per mezzo di prodotti generalmente indicati come F.T.P. (File Transfer Protocol). Questa tecnica presenta gli stessi problemi di quella precedente, a meno

dell'affidabilità, in quanto sono gli stessi prodotti che permettono il controllo della ricezione e la eventuale ritrasmissione automatica.

- Questionari : E' questa un ultima possibilità attualmente in fase di valutazione presso l'Istituto Nazionale di Statistica, mediante la quale i questionari dei censimenti potrebbero essere distribuiti e raccolti direttamente su un sito WEB della rete Internet. In questo modo il singolo cittadino potrebbe ricevere, compilare ed inviare il questionario in maniera completamente autonoma, riducendo anche in questo caso tempi e costi di trattamento ed errori di compilazione. I problemi legati a questa tecnica sono ancora una volta analoghi a quelle precedenti: autenticazione, integrità e riservatezza.

Le risposte ai problemi di sicurezza delle diverse nuove tecniche di acquisizione ci pervengono dagli strumenti che recentemente si sono resi disponibili.

E' infatti la realizzazione di una Infrastruttura a chiave pubblica (PKI) basata sulla gestione di certificati digitali X.509 lo strumento di base in grado di permettere il superamento dei problemi precedentemente evidenziati. Tramite essa infatti a ciascun utente viene assegnata una chiave di cifratura e firma, autenticata dal certificato X.509 rilasciato da un'autorità pubblicamente riconosciuta.

L'utilizzo del protocollo SSL e la cifratura fuori linea degli oggetti da spedire possono invece garantire dell'integrità e della riservatezza dei dati.

Nel caso di collegamento Extranet e quindi di possibilità di trasferimento di file di grandi dimensioni tra organizzazioni diverse si può pensare a strumenti di sicurezza che si poggiano sui livelli OSI più bassi quale ad esempio il protocollo Ipsec.

Analizzata la fase di acquisizione prendiamo in esame quella di elaborazione dei dati. Durante questa fase i problemi che si presentano sono necessariamente più diversificati, poiché ci si trova completamente all'interno del processo produttivo per il quale non solo devono essere rispettate tutte le norme sulla privacy ma deve anche essere garantita la conservazione del patrimonio informativo tanto dagli attacchi esterni quanto da quelli interni.

In questa esposizione si vogliono però mettere in risalto solo le modalità di elaborazione e i relativi problemi e le soluzioni di sicurezza che si sono venute ad affermare grazie al processo di rinnovamento di cui la produzione statistica è stata protagonista.

I più grandi fattori di cambiamento sono quelli legati al fatto di avere un'informazione non più disponibile in un unico punto, su un unico elaboratore, dove venivano concentrate tutte le politiche di sicurezza, ma ormai un'informazione disponibile in rete, accessibile da parte dei ricercatori dell'Istituto ovunque essi siano.

E' per questo, dunque, che prendono particolarmente corpo i problemi relativi, da un lato, all'autenticazione ed autorizzazione degli utenti (L'utente che vuole accedere al sistema è proprio quello che dice di essere ? E quali sono le risorse che è autorizzato ad usare ?) e dall'altro, alla riservatezza delle informazioni che transitano in rete soprattutto nel caso in cui gli utenti possano accedere da postazioni periferiche o addirittura collegarsi da postazioni mobili.

E' chiaro dunque che anche in questo caso l'esistenza di una struttura PKI può essere vista come l'elemento necessario, ma ora non sufficiente, per la garanzia dell'autenticazione. Ad essa infatti devono essere aggiunti sia strumenti sia precise procedure organizzative che permettano di gestire i profili d'utente, i gruppi e le relative autorizzazioni.

Anche per questa fase lì dove nascano problemi di riservatezza delle informazioni, a fronte di un eterogeneo utilizzo di applicazioni e strumenti, è conveniente garantire la riservatezza delle informazioni ad un livello più basso di quello applicativo e da questo indipendente con strumenti quali il protocollo Ipsec o comunque attraverso la costituzione di reti private virtuali su reti di comunicazione insicure.

La fase che maggiormente trae vantaggio dalle nuove tecniche di produzione statistica è indubbiamente la diffusione dell'informazione prodotta. Per essa si assiste ad un vero e proprio capovolgimento di approccio. Mentre prima l'informazione su supporto magnetico veniva diffusa solo come complemento alle pubblicazioni su carta, ora essa stessa è diventata informazione primaria relegando quella su carta al solo valore di copia. La nuova legislazione in materia di firma digitale ha sancito in maniera ufficiale tale capovolgimento di mentalità ed ha dato una spinta propulsiva ad organizzazioni private e pubbliche a realizzare applicazioni e servizi dirette in tal senso.

La nuova diffusione di dati statistici quindi si avvale dei seguenti strumenti:

- WEB: Si è ormai affermato come uno degli strumenti principali di diffusione, evolvendosi da una pubblicazione completamente statica delle informazioni (è il fornitore di informazioni che decide quali dati presentare ed in quale maniera) ad una informazione dinamica (è l'utente che interagendo con i database decide quali dati estrarre ed in che maniera organizzarli). Il salto di qualità è enorme e richiede un ripensamento delle politiche di sicurezza nei confronti degli accessi dall'esterno. In questa ottica infatti il fornitore di informazioni mette a disposizione della comunità scientifica il proprio patrimonio informativo. E' necessario quindi poter autenticare gli utenti e concedere ad essi diversi livelli di autorizzazione. Di fondamentale importanza è poi il problema dell'integrità non solo dei sistemi WEB server

ma anche degli stessi contenuti delle informazioni pubblicate, soprattutto nel caso in cui queste riguardino dati statistici.

- Applet : Direttamente legata alla diffusione via WEB è quella di poter costruire interfacce sempre più complesse e performanti che permettano una navigazione all'interno dei sistemi sempre più vicina alle esigenze degli utenti. Tali obiettivi possono essere raggiunti attraverso la realizzazione di "Applet" da scaricare direttamente sulla stazione client dell'utente. Questa soluzione è però perseguibile solamente se si riesce a garantire l'utente stesso della origine non dolosa dell'applicazione.
- Documenti legali: All'interno dei processi di diffusione statistica esiste anche la possibilità di diffondere a richiesta certificati legali relativi a particolari situazioni (si pensi ad esempio al valore dell'indice dei prezzi al consumo cui sono legati alcuni contratti). La nuova normativa relativa alla firma digitale permette di pensare a tali certificazioni in forma completamente digitale, in cui l'eventuale versione su carta ne diventa solamente una copia. E' chiaro in questo caso che i più rilevanti problemi di sicurezza oltre a quelli relativi all'integrità nella trasmissione, si rivolgono soprattutto al rispetto delle norme che garantiscono la legalità del documento elettronico.
- E-book: La nuova frontiera della diffusione delle pubblicazioni statistiche è quella rappresentata dalla editoria elettronica, per mezzo della quale i libri, le pubblicazioni vengono rese disponibili per l'acquisto direttamente in rete. Le problematiche che si vengono a creare in questo caso sono di ordine completamente diverso in quanto si riferiscono soprattutto alla protezione degli interessi dell'organizzazione in termini di copyright.

Le tecniche di sicurezza che sovrintendono ai problemi segnalati nei punti precedenti, oltre ad una diversa progettazione dei sistemi di firewalling che permettono di proteggere le risorse interne da attacchi provenienti dalla rete , si possono ricondurre ancora una volta alla realizzazione di strutture PKI ed alla applicazione della firma digitale. Quest'ultima in particolare permette di autenticare contenuti, applet, documenti legali, licenze di e-book. Essa quindi può essere la giusta risposta alle esigenze di sicurezza legate alla fase di diffusione.

Ma è giusto chiedersi quali implicazioni possa comportare un tale quadro di rinnovamento. Per semplicità concentriamo l'attenzione solo su quella che sembra essere una soluzione ad ampio spettro in grado di coprire le esigenze di sicurezza nelle tre fasi principali del processo di

produzione statistica ma che al tempo stesso costituisce il vero elemento innovativo all'interno dell'organizzazione: la realizzazione di una PKI.

Questa richiede una struttura organizzativa interna, inizialmente non prevista, che sia in grado di gestire completamente il processo di rilascio e revoca delle chiavi e dei relativi certificati, all'interno di un processo di sicurezza delineato nei minimi dettagli. La costituzione di una tale struttura si dovrà avvalere di nuove figure professionali che è necessario formare su tematiche ancora non note alla maggior parte e che spesso appaiono addirittura "misteriose" ai non addetti ai lavori.

Questa sensazione è tanto più viva quanto più l'ambiente risulta non essere completamente sensibile alle problematiche di sicurezza, in cui gli utenti in alcuni casi evitano addirittura di ottemperare agli obblighi imposti. Risulta perciò necessario nel momento in cui si affrontano questi nuovi scenari introdurre nell'organizzazione e a tutti i livelli una nuova cultura che vede la sicurezza come un bene primario irrinunciabile.

Ma la definizione di una struttura PKI non è solo organizzazione, ma richiede anche la definizione di alcune scelte di base di non facile soluzione. Su che cosa investire dunque? Sulla realizzazione di una struttura completamente in azienda, struttura di hardware, software, sicurezza e persone capace di mantenere il servizio costantemente attivo e di interagire con gli utenti, anche in presenza di numeri che possono rapidamente crescere in funzione delle esigenze di comunicazione e di relazione proprie dell'azienda ? Oppure conviene rilasciare tutto ciò in "outsourcing", demandando completamente tutte le problematiche ad un partner fidato, ma a quali costi ? Oppure ancora optare per l' "housing" delle macchine, mantenendo in proprio la responsabilità della gestione dei rapporti con gli utenti ? E' su queste problematiche e sulla scelta più appropriata a seconda delle esigenze e dei costi preventivati che si gioca la buona riuscita di un'operazione che può rappresentare un decisivo salto di qualità del processo di produzione.

L'ultimo punto di considerazione è quello relativo alla definizione di una chiara politica nei confronti degli utenti e dei partner. Nella definizione di una PKI il passo più ovvio è rappresentato dalla distribuzione di proprie chiavi a tutti gli utenti e partner. Ma è questa una giusta politica ? Con il diffondersi delle tecniche di firma digitale potrà non essere lontano il momento in cui il generico utente dovrà possedere tante smart-card quanti sono i sistemi cui abbia diritto di accedere.

Ed allora appare giusta la preoccupazione di vedere ancora non attuata la così detta "Cross-certification", ossia la possibilità che un'autorità di certificazione riconosca i certificati digitali X.509 rilasciati da un'altra autorità. Sarebbe infatti questa la maniera ottimale di gestire l'intera utenza della propria azienda, senza dover necessariamente fornire chiavi e certificati a chi già ne è in possesso.

Al di là comunque delle implicazioni messe in luce è importante considerare come applicazioni di strumenti di questo tipo possano portare benefici all'intero processo di produzione in termini di snellimento delle procedure, maggiore tempestività nel rilascio dei risultati, maggiore accuratezza nell'acquisizione dei dati ed estensione dell'utenza a valori fino a pochi anni fa completamente impensabili.

In conclusione l'adozione di tali strumenti ed il ripensamento delle modalità di produzione statistica soprattutto nelle fasi di acquisizione e diffusione portano ad un sensibile miglioramento del servizio reso al paese.

E' tempo, dunque, che la sicurezza non vada più vista come " una tassa da pagare in termini economici, prestazionali ed organizzativi", come alcune volte ancora accade, ma come un investimento ormai irrinunciabile, capace non solo di garantire il patrimonio delle aziende e di preservarne l'immagine nei confronti dell'opinione pubblica ma anche di rappresentare uno strumento indispensabile per cogliere le nuove opportunità offerte dal mercato. Essa va quindi intesa come centro di profitto e non più come centro di costo.

Quinta
Conferenza
Nazionale
di Statistica

WORKSHOP

Editoria elettronica per la diffusione dell'informazione statistica

coordinatore

Donato Speroni
direttore di Documenti.net

Roma 15, 16, 17 novembre 2000
Palazzo dei Congressi
Piazzale Kennedy - EUR

ISTAT

Editoria elettronica per la diffusione dell'informazione statistica

coordinatore
Donato Speroni
direttore di Documenti.net

SOMMARIO

G. Carchesio, G. Mottura	E-book: il passaggio al digitale
A. Braghin e P. Crivelli	La realizzazione di “e-book” con l’utilizzo di acrobat
M. Giacobbe	Editoria elettronica vs. cartacea: caratteristiche a confronto
V. Patruno, M. Venturi	La popolazione residente nei comuni italiani

WORKSHOP

Editoria elettronica per la diffusione dell'informazione statistica

E-book: il passaggio al digitale

G. Carchesio, G. Mottura

E-book: il passaggio al digitale

Giovanni Carchesio, Giulia Mottura
Istat – Dipartimento Diffusione e Banche Dati

{carchesi,mottura}@istat.it

Abstract
Quale soluzione tecnologica risponde pienamente alle necessità e alle problematiche poste dalla scelta di pubblicare in forma elettronica? Come soddisfare l'elemento di personalizzazione che sottende a tale scelta?
Obiettivo dell'intervento è quello di proporre un contesto tecnologico che risponda a tali quesiti e renda attuabile, per un ente della tipologia a cui appartiene l'Istituto Nazionale di Statistica, il cruciale passaggio dall'editoria cartacea a quella digitale, dando così realizzazione al principio *"electronic dissemination first"*.
E' prevedibile che la consistente riduzione dei costi, legati alla tradizionale catena di produzione di documenti cartacei, avvenga parallelamente all'ottimizzazione delle fasi di distribuzione: la peculiarità dell'acquisto di prodotti digitali risiede infatti nella possibilità di essere negoziato ed effettuato su Internet; anche il relativo recapito può avvenire tramite la rete.
In breve tempo, molto materiale esistente verrà convertito in formato digitale, trasformando prodotti costituiti da atomi in prodotti costituiti da bit. Questa trasformazione è applicabile anche alla *proprietà intellettuale* (*copyright*) relativa al medesimo materiale?
La risposta a tale quesito deve essere ricercata nelle nuove tecnologie riunite sotto l'acronimo D.R.M. (*Digital Rights Management*): il DRM permetterà infatti ad editori e autori di proteggere le loro proprietà intellettuali dall'utilizzo e dalla distribuzione illegale. Tali tecnologie agevoleranno i distributori di prodotti digitali nel *controllo* o nella *limitazione dell'accesso* agli stessi prodotti.
Mediante una sorta di *codifica personalizzata* realizzata tramite crittografia e firma digitale, verranno infatti fornite agli utenti finali delle *chiavi personali* per la visualizzazione dei prodotti, con possibili limitazioni sulle funzionalità di copia, stampa, estrazione ed eventuale ridistribuzione.
Rispetto alle peculiarità delle pubblicazioni Istat, il formato digitale permette l'inclusione e l'estrazione di "formati nativi". Ad esempio, sarà possibile allegare ad un libro elettronico – o *eBook* – dati statistici (tabelle) ma anche software di elaborazione degli stessi (file eseguibile). Il prodotto finale consiste in un unico "pacchetto" digitale (*envelope*) tale da poter essere distribuito tramite Internet.
L'utente potrà scegliere se estrarre un intero documento oppure parte di esso (ad esempio i soli dati in modalità nativa), ma anche se stamparlo solamente o copiarlo per eseguire rielaborazioni. Il fondamentale vantaggio che l'ente produttore di informazioni offre infatti alla propria utenza, tramite l'adozione di questa soluzione, consiste proprio nel dare l'opportunità al singolo destinatario di personalizzare le caratteristiche d'uso del prodotto da acquisire.

QUINTA CONFERENZA NAZIONALE DI STATISTICA
Innovazione tecnologica e informazione statistica
"Editoria elettronica per la diffusione dell'informazione statistica"
ROMA 15-16-17 novembre 2000
Palazzo dei Congressi - EUR

La comunicazione e la diffusione tramite il web: il caso dell'Istituto nazionale di statistica nell'ambito della P.A.

Una delle grandi potenzialità del mezzo Internet, rispetto alla stampa e ad altri media, sta nel suo carattere potenzialmente esaustivo, dovuto alla relativa assenza di limiti di spazio e di tempo.
L'utente del web ha quindi a sua disposizione diversi livelli di approfondimento e diversi generi di informazione: dalla notizia flash di agenzia di stampa all'ampio servizio monografico, dal commento all'analisi; inoltre ha la possibilità di consultare agevolmente schede, grafici, tabelle, ma anche documenti originali. Vantaggio principale è che a tale mole di informazioni si possa accedere direttamente, senza alcuna mediazione. Internet diventa quindi soprattutto una fonte di primaria importanza, oltre che strumento di ricerca.
La fine dell'intermediazione (ossia l'opportunità di attingere direttamente alla fonte) si può considerare, infatti, una delle principali peculiarità dello scambio in senso lato, attraverso la rete. Dall'acquisto di un libro o di un cd-rom, al download di dati da un database, è possibile per l'utente confrontare direttamente una pluralità di offerte, scegliere ciò che più risponde alle proprie esigenze e, nel caso di transazioni di natura commerciale, ordinare immediatamente ciò che si è scelto. Software specializzati[1] possono, anche se per ora in modo poco soddisfacente, aiutare l'utente nella ricerca, raccogliendo le proposte selezionate in base ad una richiesta mirata.
Un costante stato di mutazione caratterizza Internet a causa dell'ingresso in campo di sempre nuovi soggetti, sia come promotori di siti, sia come fruitori-utenti.
Inoltre il Web diventa sempre più il terreno di scontro e di un rinnovato incontro tra comunicazione testuale e comunicazione sonora e visiva.
Le novità tecnologiche contribuiranno a migliorare efficienza e fruibilità del web e a colmarne gli attuali limiti, costituiti principalmente dal linguaggio di rappresentazione utilizzato (HTML). Le pesanti restrizioni imposte all'autore della pagina web, soprattutto riguardo al valore semantico da attribuire agli elementi che la costituiscono, appaiono ancora piuttosto evidenti. Tale ostacolo potrà essere superato dall'introduzione di un nuovo linguaggio (XML), dotato di caratteristiche tecniche che permettono il superamento dei limiti dell'HTML, riguardanti soprattutto l'indicazione del significato e del valore contenutistico di parole e immagini.
E' chiaro comunque che la rete deve essere considerata luogo privilegiato per la diffusione, la comunicazione e lo scambio di dati, notizie, informazioni dei generi più disparati.
Tale affermazione è valida per tutti i soggetti attivi nella società contemporanea. Tanto più per gli enti pubblici, che svolgono, secondo la loro missione, un'attività rivolta al servizio per la cittadinanza.
Adeguare le pubbliche amministrazioni alle esigenze di una moderna società è ormai, in effetti, un obiettivo non procrastinabile, se si vuole un'amministrazione che produca servizi migliori, con costi sempre minori per la collettività[2].
Compito dei governi deve essere quello di promuovere ogni azione utile ad accrescere la diffusione della cultura e delle tecnologie dedicate alla comunicazione e

[1] I motori di ricerca sono programmi raggiungibili attraverso la rete, in grado di indicizzare e fornire, su richiesta dell'utente, informazioni presenti su Internet. Difficile è poi, per l'utente, una volta reperita l'informazione cercata, valutarne la sua correttezza ed imparzialità.
[2] La Società dell'informazione ha ridefinito il concetto di partecipazione, e quindi di cittadinanza. Si tratta di uno dei cardini fondamentali delle società contemporanee, che oggi deve comportare il diritto per tutti di usufruire di nuovi servizi online. Lo stesso concetto di servizio universale si è trasformato, essendo ormai impossibile limitarlo alla sola connessione alle linee telefoniche. E' necessario quindi ampliarlo, fino ad includere la possibilità di accesso ad una struttura di telecomunicazione veloce e sicura, e ai servizi che in essa transitano.

all'informatizzazione nelle pubbliche amministrazioni (anche a livello locale e per tutti gli enti che dipendono dallo Stato).
Tale attività potrà favorire sia un rapido aumento dei livelli di soddisfazione dell'utenza, sia una maggiore efficienza del sistema amministrativo che agevoli anche la competitività globale dell'intero Paese.
Questi obiettivi saranno raggiunti solo attraverso la valorizzazione e la diffusione delle esperienze realizzate in materia di nuovi servizi e la revisione di quelli consolidati.
L'evoluzione del settore della comunicazione e, in prospettiva, la diffusione delle reti a larga banda che permetteranno l'accesso ai servizi multimediali e interattivi, ha posto al centro del dibattito politico-istituzionale, a livello nazionale ed europeo, l'evoluzione del concetto di servizio universale[3]. L'uso delle nuove tecnologie informatiche e telematiche, deve essere riferito alla generalità dei cittadini, e in tal senso, la natura dei nuovi servizi si dovrebbe connotare per il carattere dell'universalità.
In Italia è attivo un "Forum per la Società dell'Informazione", promosso dalla Presidenza del Consiglio dei Ministri. Tale Forum, che si muove in sintonia con il "V programma quadro di ricerca e sviluppo 1998-2002" dell'Unione Europea, è aperto alle istituzioni pubbliche, anche territoriali, alle università, agli istituti di ricerca, alle parti sociali, ai cittadini. Lo scopo dell'iniziativa è quello di elaborare proposte che costituiranno la base del Piano di Azione dell'Italia per lo sviluppo della Società dell'Informazione; ciò al livello, oltre che di pubblica amministrazione, anche di infrastrutture e reti tecnologiche; occupazione, tecnologie e organizzazione industriale, forme di lavoro e figure professionali; alfabetizzazione informatica, formazione, tecnologie della didattica e ricerca; servizi e contenuti multimediali[4].
Anche l'Istituto nazionale di statistica è coinvolto in questo cambiamento epocale, che sta portando mutamenti in ogni ambito della sua organizzazione.
In particolare, l'Istat, negli ultimi anni, ha sostanzialmente modificato la propria politica di comunicazione e diffusione delle informazioni prodotte.
Tale cambiamento si è realizzato non solo nei confronti dei media, degli operatori economici, degli opinion leader, ma anche dell'opinione pubblica in generale.
In particolare, l'Istat ha scelto di passare dalla sola produzione di volumi cartacei alla diffusione online (web e banche dati) e offline (dischetti e cd-rom). Tale scelta è stata dettata dalla volontà di raggiungere una maggiore tempestività nel rilascio dei dati e per consentire una più efficiente organizzazione della grande massa di dati disponibili.
Le iniziative, volte a facilitare l'accesso alle informazioni di natura statistica da parte di una variegata utenza, consentono un costante aggiornamento sui principali dati statistici, sulle novità editoriali, sull'attività in generale dell'Istat. Ciò è possibile, oltre che tramite la pubblicazione di newsletter e notiziari, soprattutto attraverso:

- il **sito web** dell'Istituto (www.istat.it);
- il **B.B.S.** (Bulletin Board System)[5], un archivio elettronico online contenente le pubblicazioni dell'istituto (bbs.istat.it);
- le **banche dati**, consultabili presso i Centri di Informazione Statistica regionali, le Camere di Commercio ed anche tramite Web (come nel caso della banca dati sul Censimento Intermedio dell'Industria e dei Servizi, di "Demo" la banca dati sulla popolazione italiana, di "ConIstat" la banca dati dedicata alla congiuntura economica).

[3] Il servizio universale inteso come accesso ad un servizio di telecomunicazioni, è considerato come un diritto fondamentale di ogni cittadino, essenziale per la piena appartenenza di un individuo ad una comunità e come elemento costitutivo del diritto alle libertà di espressione e comunicazione.
[4] Lo sviluppo di Internet e dell'Industria dei servizi sono elementi fondamentali per la crescita del Paese. Occorre creare le premesse affinchè si realizzi un circolo virtuoso in cui servizi multimediali e infrastrutture evolvano in modo integrato.
[5] Il BBS Istat, nato come un sistema telematico costituente una sorta di bacheca elettronica, è ad oggi un vero e proprio sito web. In esso vengono raccolti files di dati scaricabili in linea.

Oltre agli esperti che accedono ai dati dell'Istat, a chi richiede elaborazioni specifiche a pagamento e agli utenti che leggono le statistiche sui giornali, emerge un ulteriore tipo di utente, che non reputa sufficiente l'informazione fornita da giornali e televisione, ma vuole invece potersi documentare in modo diretto e interattivo, attraverso moderni sistemi di informazione, consultando direttamente i dati alla fonte.
I dati statistici trattati dalla stampa e dai media in genere, sono infatti quantitativamente molto inferiori rispetto alle statistiche realmente disponibili. Tale paradosso risulta ancor più evidente se si pensa alle statistiche riferite ai temi sociali, che solo ultimamente stanno emergendo agli occhi dei media, ma che generalmente riscuotono minore attenzione rispetto, ad esempio, alle statistiche economiche.
In realtà ci si avvicina gradualmente al momento in cui anche i singoli cittadini potranno usufruire dell'informazione statistica a diversi livelli: attraverso i media tradizionali, ma anche attraverso le reti informatiche, i cd-rom, i sistemi multimediali, tramite i quali si potranno recuperare tutte le informazioni più o meno dettagliate, nel momento in cui l'utente è interessato ad una lettura e ad una analisi meno frettolosa dei dati. La lettura più approfondita dei dati può ovviamente essere realizzata tramite supporto cartaceo, ma è forte la differenza tra la ricerca in una biblioteca e la possibilità di arrivare agli stessi dati, ad esempio, navigando su Internet.
Per rispondere a tali esigenze l'Istat ha messo a disposizione dell'utenza, già da alcuni anni, il proprio Bulletin Board System (BBS), dal quale è possibile scaricare novità editoriali e altre informazioni di base. Tale strumento, per molto tempo ad accesso gratuito, ma recentemente aperto solo agli abbonati, è da considerare come un antenato delle banche dati[6] consultabili interattivamente.
Il ruolo e le funzioni del BBS sono quindi da rivedere, anche in base alla evoluzione del sistema di banche dati e all'integrazione con il sito web dell'istituto.
Recentissima è, inoltre, la disponibilità per l'utenza online di ConIstat, la banca dati nata come cd-rom delle statistiche congiunturali. Tale prodotto ha riscosso un notevole successo di pubblico, essendo un vero e proprio sottosistema informativo di tutte le statistiche congiunturali, costantemente ampliato ed integrato, ma soprattutto interrogabile in maniera interattiva.

Un sistema integrato di diffusione online per l'Istat

In effetti nel nostro Paese, domanda ed offerta di dati statistici sono cresciute in modo esponenziale, negli ultimi anni. L'Istat, con i suoi comunicati, i volumi, le banche dati, è certamente il principale produttore in questo settore.
Il confine tra informazione gratuita e a pagamento è però molto difficile da tracciare. Se infatti, da un lato, si reputa realizzabile la possibilità, per un ente come l'Istat, di autofinanziare almeno in parte le proprie attività, dall'altro non è pensabile che l'Istituto deroghi alle proprie funzioni di diffusione della conoscenza e della cultura statistica, anche quando tali attività non dovessero rivelarsi remunerative.
E' sempre più urgente che sia definita con chiarezza, la distinzione tra prodotti quasi esclusivamente gratuiti e prodotti a pagamento, destinati al vasto pubblico.
E' probabile che il problema di cosa far pagare agli utenti trovi una sua naturale soluzione ed una logica risposta nell'omogeneizzazione delle politiche sugli standard proposti da Eurostat, a livello comunitario.

[6] A livello europeo, milioni di dati statistici, comparabili a livello comunitario, sono consultabili attraverso le banche dati. "New Cronos" contiene informazioni statistiche sui più importanti aspetti degli stati membri dell'Unione europea, dei paesi dell'EFTA, degli Usa e del Giappone. "Comext" fornisce dati sul commercio con l'estero degli Stati membri per oltre 10.000 prodotti. Tali banche dati però, non sono consultabili online, ma solo presso i data shop Eurostat di Roma e di Milano. In tali spazi, sono disponibili anche pubblicazioni e prodotti offline.

In linea di massima, la soluzione corretta sembra essere quella di far pagare il materiale di approfondimento pubblicato "al costo", tenendo presente che in tale costo siano inclusi, oltre che stampa, carta e supporto informatico, anche i costi di organizzazione e distribuzione finale del materiale.
Tutte le pubblicazioni a pagamento dovrebbero essere acquisibili attraverso tecniche di commercio elettronico, sia che si tratti di dati online, sia di altro tipo di materiale immesso sul mercato dall'Istituto, che deve raggiungere l'utenza-clientela nel minor tempo possibile. Non a caso, la scelta su cui si stanno orientando tutti gli Istituti nazionali di statistica si può condensare nel principio "*Electronic First*".
Per l'Istat essa comporta una serie di implicazioni: la creazione di uno o più data warehouse, lo sviluppo dell'offerta su Internet, la modificazione della politica di comunicazione e diffusione delle informazioni prodotte.
L'architettura generale del sistema di diffusione in rete e su supporto informatico dei dati statistici, finalizzato anche all'agevolazione dell'accesso rapido dell'utenza ai dati in linea, necessita di un elevato grado di flessibilità che permetta anche di integrare moduli differenti, realizzati in futuro con l'apporto di tecnologie emergenti.
Fondamentale è quindi la realizzazione di applicazioni di diffusione che permettano, ad esempio, di attingere ai processi produttivi messi a punto dai servizi, superando la necessita di intervenire sui dati con processi di assemblaggio e nuova formattazione. Tale sistema deve essere caratterizzato, soprattutto dal punto di vista tecnico, da elevata performance ed affidabilità.
Data ormai per certa, quindi, la continua crescita dell'informazione statistica da immettere in linea - tenendo anche conto dell'impegno che comporterà la diffusione dei censimenti del 2000-2001 (popolazione, agricoltura, industria e servizi) - è importante per l'Istat superare la filosofia del singolo giacimento di dati onnicomprensivo. Più adatto alle esigenze dell'Istituto sembra, infatti, essere un sistema integrato di diffusione online con interfacce e protocolli Internet in cui si possa realizzare "web farming"[7].
La modalità principale di trasferimento delle informazioni statistiche è infatti il web. Ed il sistema deve essere concepito per interfacciare direttamente l'utenza esterna che lo consulta via web e gli operatori dell'Istituto fornitori delle informazioni e dei dati.
Le applicazioni del sistema devono permettere inoltre, sempre senza interventi diretti sui dati, l'estrazione di files ad hoc per l'utenza, di file per la stampa dei volumi e la generazione di floppy e cd-rom.
E' possibile considerare, all'interno del sistema, la conservazione del BBS come strumento di scelta e trasferimento di file preconfezionati via Internet (dei quali è però importante garantire l'uniformità di formato).
L'architettura del sistema deve prevedere l'accesso dell'utenza tramite il sito web istituzionale, che ha un ruolo di presentazione dell'Istat; questo deve contenere, oltre ai meccanismi di instradamento, anche quelli di riconoscimento ed accreditamento dell'utenza. L'utente viene quindi indirizzato ai vari elementi che compongono il sistema e alle applicazioni di suo interesse. Ciascuna applicazione potrebbe essere

[7] Il termine "web farming" trae origine dall'incontro tra data warehousing e tecnologie web. Il paradigma che il web sottintende è radicalmente differente dal paradigma insito in un data warehouse. Il contenuto web può riferire qualsiasi cosa (spaghetti data), è altamente volatile e in continuo mutamento. Spesso invece occorre operare in un ambito disciplinato per trasformare dati grezzi in informazioni valide.
Evitando di navigare sul web in maniera caotica, passando da un item informativo all'altro, il web farming si concentra su un processo evolutivo e sistematico di ricerca, acquisizione, strutturazione, raffinamento e disseminazione di contenuti ("Information Refining").
Un sistema di web farming deve confezionare e recapitare (disseminazione), alle persone giuste e al momento giusto, informazioni rilevanti ed appropriate, tenendo in debito conto questioni importanti come privacy, confidenzialità e proprietà intellettuale. Vengono adottati vari meccanismi di disseminazione, da attività schedulate e determinate a priori fino a query ad hoc.

considerata e realizzata come un web autonomo, con un proprio indirizzo ma legata strettamente al sistema con gli opportuni link.
L'efficace gestione del sistema ed il suo corretto sviluppo sono comunque subordinati all'affermarsi di una cultura che ponga l'utente e le sue esigenze al centro del processo di diffusione dell'informazione statistica e soprattutto che non veda l'utenza come possibile pericolo per il sistema.
La commercializzazione dei prodotti dell'Istat (editoria, banche dati) trova nella rete un passaggio ormai inevitabile. Non si può non tenere conto quindi, dello spazio che tale aspetto e tali attività debbano occupare sul sito web.
E' da non escludere a priori addirittura che in futuro il sito non possa accogliere sponsor e pubblicità anche di terzi (discorso forse ancora prematuro per un ente pubblico come l'Istat).
In ogni caso è da definire una strategia di accreditamento e pubblicità sui più noti e frequentati motori di ricerca esterni.
Sempre ai fini della commercializzazione dei prodotti Istat, è da reputare inoltre necessaria la predisposizione nella struttura del sito, ma anche nelle applicazioni, l'implementazione delle funzionalità inerenti al commercio elettronico di tipo "one stop shop", per la realizzazione di un'edicola elettronica delle informazioni statistiche.
I servizi commerciali offerti all'utenza devono infatti essere sostenuti da funzionalità di base quali: la gestione dei pagamenti attraverso carte di credito e prepagate, della sicurezza dei trasferimenti, della certificazione ed autenticazione mediante tecniche di crittografia a chiave pubblica, dei servizi d'assistenza ed interattivi al pubblico (per esempio elaborazioni su commissione).
I servizi a pagamento richiedono ovviamente una risposta puntuale e precisa; all'utente, soprattutto a quello pagante (tale principio dovrebbe in realtà essere valido per tutta l'utenza: anche per i servizi online accessibili gratuitamente), bisogna quindi assicurare una risposta certa e affidabile.
Per realizzare tale obiettivo, è necessaria la razionalizzazione di efficienti data entry operativi: tutto ciò soprattutto per la gestione dei cataloghi elettronici, degli archivi tradizionali ed ipermediali, oltre che delle informazioni "real time" (comunicati e note per la stampa, eventi, novità).
Allo scopo di apportare sempre nuovi miglioramenti al sito web è bene garantirsi la possibilità di tenere il bacino d'utenza del sito sotto continuo o periodico monitoraggio, attraverso, ad esempio, questionari online, gestione delle abilitazioni e password per gli accessi, collegamenti con sistemi informativi di marketing.
Inoltre, per adeguare in modo sempre più efficace la propria offerta alla domanda della variegata utenza, l'Istat è al momento impegnato nella ricerca di innovative modalità di diffusione. I risultati di recenti indagini hanno infatti consolidato la consapevolezza che le possibilità offerte dalle nuove tecnologie siano ben gradite a tutti coloro che si rivolgono all'istituto per riceverne informazioni e dati.
Anche a livello internazionale la strada ormai perseguita è quella della diffusione tramite le vie di comunicazione di tipo digitale.
L'editoria elettronica, in particolare, può rappresentare il naturale sviluppo dei tradizionali sistemi usati dall'Istat per la diffusione dei risultati delle proprie ricerche: dalla scelta digitale scaturirebbe la necessità di una totale ridefinizione dei canali informativi destinati all'utenza, un ripensamento dell'uso del supporto cartaceo, da integrare con la produzione su supporto elettronico.

Editoria elettronica e gestione dei diritti digitali

La diffusione dell'informazione statistica è stata notevolmente modificata da Internet. I tradizionali canali di diffusione e, in particolare quello dei volumi cartacei, sono allo stato oggetto di evoluzione. Dando ad ogni utente l'opportunità di servirsi

direttamente online, le necessità di arrivare alla fase della stampa sembrano ridursi progressivamente; in special modo se la trasposizione online del volume consiste in una omogenea raccolta di dati organizzati in tabelle facilmente indicizzabili e quindi consultabili separatamente. Nel caso in cui il volume consista in un prodotto più eterogeneo (gli *annuari statistici*) e la consultazione su carta è difficilmente sostituibile, la realizzazione della stampa del volume risulta ancora preferibile.
I monitoraggi condotti presso i C.I.S. (*Centri di Informazione Statistica*), dal 1996 a oggi, rivelano che il 66% degli utenti riceverebbe volentieri informazioni mediante la posta elettronica (o comunque su Internet), anche se permangono dubbi sui formati che tali informazioni dovrebbero assumere.
Tutto ciò che attualmente viene rilasciato, scaricato e immesso in Internet da parte dell'Istat può infatti avere diversi formati. Tali formati vengono generati da sistemi dinamici di interrogazione che estraggono informazioni (*data warehouse*) e, nel caso dei dati statistici, producono file ASCII, tabelle EXCEL ma anche formati più "eleganti" quali PDF e HTML. In generale si assiste alla personalizzazione dell'accesso verso una grande massa di dati, potendo nel contempo selezionare e costruire rapidamente le informazioni utili a ciascun utente.
Accettare questa logica è necessario ma tutt'altro che semplice, sia per problemi di risorse ma anche perché essa implica un ripensamento dei processi di produzione. La struttura del prodotto/servizio dovrebbe essere definita sin dall'inizio, e non, come oggi avviene, a metà o verso la fine del processo di raccolta e produzione delle informazioni statistiche. Si prospettano quindi, per rispondere alla citata necessità di continuo miglioramento dei prodotti e dei servizi, nuovi investimenti, risorse dedicate e cambiamenti organizzativi. Gli istituti di statistica nazionali, Eurostat e molti enti Sistan stanno affrontando queste problematiche, sempre più legate ai costi di produzione.
La *regola aurea* degli istituti di statistica è stata, ad oggi, quella di far pagare agli utenti il costo marginale della produzione dei supporti e dell'erogazione dei servizi. I prezzi per recuperare le spese editoriali e di diffusione sono sempre stati piuttosto variabili, più alti in Europa del Nord (il prezzo include tutte le fasi che seguono la rilevazione e l'elaborazione) e inferiori nei paesi del Mediterraneo (il prezzo si identifica con il costo della tipografia). Seppure declinata vari modi, tale regola è stata valida per tutti gli istituti di statistica e per Eurostat fino all'avvento di Internet.
Il costo di un libro cartaceo attualmente si può ripartire come segue:

- stampa (10%);
- gestione (20%);
- tasse e autori (10%);
- distribuzione e promozione (60%).

Nell'ambito dell'attività di diffusione dei vari istituti di statistica europei si è comunque rilevata una tendenza sempre più marcata a rendere i dati disponibili in rete e stampare soltanto quanto è strettamente necessario, riducendo la prima delle voci di costo citate. Per poter agire anche sull'ultima di tali voci, non rimane che intervenire sulla distribuzione ovvero rendere i prodotti disponibili sulla rete ed evitare, o perlomeno contenere, il problema della distribuzione. Bisogna quindi poter rendere effettivo lo slogan "*electronic first*" per tutta la produzione editoriale.
Gli utenti devono supportare i costi di tale operazione? Se la risposta a tale quesito è negativa, gli stessi costi dovrebbero essere coperti dallo Stato (trattandosi di servizi fondamentali per il paese) con finanziamenti dedicati oppure presi in carico dal settore privato attraverso il sistema della pubblicità.
A tale proposito, di recente è stato lanciato un sistema per inglobare pubblicità all'interno di brani musicali (ma anche video, libri, giochi e altre forme di distribuzione digitale dei contenuti). I file corrispondenti vengono etichettati con brevi informazioni commerciali (anche audio e video). In questo modo, da un lato i brani potrebbero

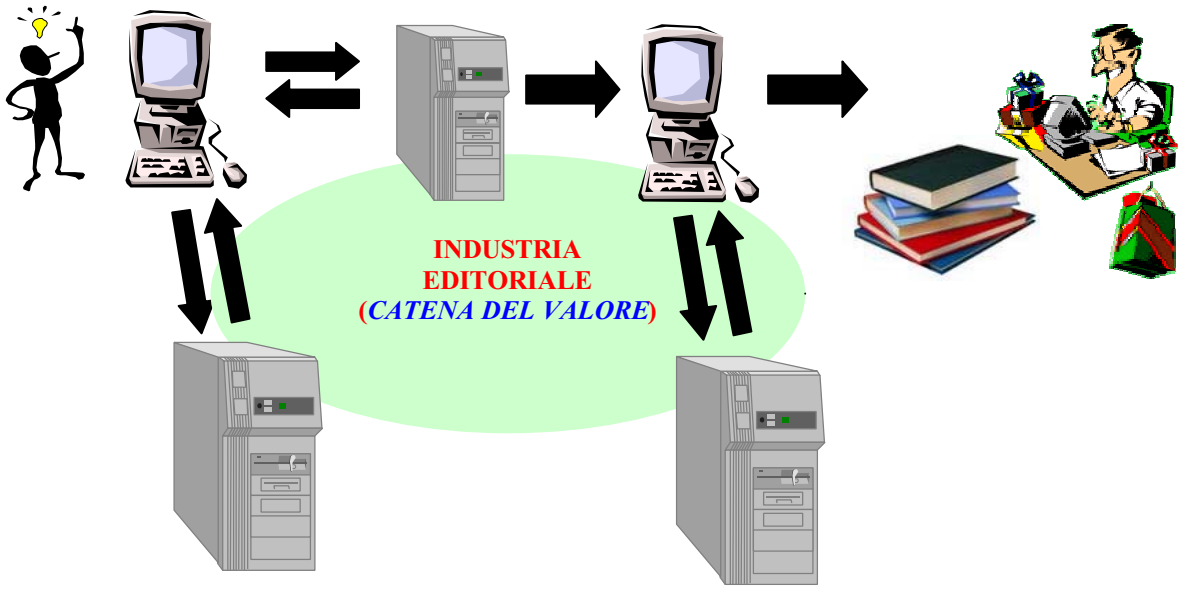
essere utilizzati gratuitamente dagli utenti finali e, dall'altro, gli autori (e chiunque detenga diritti economici) sarebbero ripagati grazie ai proventi pubblicitari.
Anche volendo dare per ormai acquisito il principio di "*electronic dissemination first*" (prima si rilascia il dato online, poi su carta) restano aperti vari interrogativi. Quanto deve essere ridimensionata l'editoria tradizionale nella scelta dei testi e nelle tirature? Quanto è davvero vicina l'ipotesi dei libri on demand? Come tutelare il copyright senza imporre forme troppo restrittive, incompatibili col dovere di diffusione proprio della statistica pubblica? Bisogna affrontare il confine tra informazioni gratuite e a pagamento, confine difficilissimo per la statistica ufficiale, così come è necessario arrivare ad una realistica valutazione del commercio elettronico nella diffusione dei dati statistici o nella vendita di pubblicazioni.
Il nuovo millennio porta con sé un'ulteriore trasformazione: il "*libro virtuale*" o "*eBook*", un oggetto che dopo molti secoli abbandona il supporto cartaceo come connotato primario ed entra nel mondo dei bit (pura informazione). Bisogna analizzare la portata di questa trasformazione nei suoi aspetti innanzitutto tecnologici, ma anche per le sue implicazioni in una politica editoriale allargata e incentrata sul web.

Evoluzione dell'editoria elettronica (Digital Publishing)

Per una migliore comprensione dell'attuale *confluenza tra web ed editoria*, risulta utile proporre una interpretazione, già abbastanza diffusa, degli *stadi di sviluppo* attraverso i quali l'editoria elettronica si è evoluta e probabilmente continuerà ad evolvere. Vengono individuati "**5 stadi evolutivi**", caratterizzati singolarmente in base alla tipologia di prodotti editoriali, ai formati, alle tecnologie di stampa e all'eventuale utilizzo di supporti informatici e telematici.
Lo "**scenario 0**" è quello caratterizzato dall'**editoria convenzionale** (dal 1450), iniziato con la rivoluzione provocata da Gutenberg. L'editoria offre contenuti informativi mediante il prodotto a stampa ovvero la classica pubblicazione cartacea. Molte copie dello stesso prodotto vengono stampate per trarre vantaggio dalle economie di scala realizzabili sul processo di stampa. Il punto iniziale della catena editoriale è costituito dai *contenuti originali* dell'autore, disponibili in forma scritta o dattiloscritta. Tali contenuti poi vengono trasformati in pubblicazioni facendo ricorso a un formato di stampa.

Lo "**scenario 1**" introduce le funzionalità di **supporto elettronico all'editoria convenzionale** (dal 1985) mediante l'adozione di nuove tecnologie e tecniche compatibili con l'editoria convenzionale. Ha inizio con l'utilizzo di cd-rom come mezzo supplementare di editoria e l'uso di posta elettronica per la comunicazione tra i vari ruoli editoriali. Vengono realizzate e usate reti (network) per il trasferimento di materiali tra i vari ruoli editoriali. Dal punto di vista organizzativo si assiste ad una progressiva concentrazione di compiti diversi in un singolo ruolo altamente specializzato. In alcuni casi tale *specializzazione* permette di ridurre i costi dell'editore (autori che realizzano illustrazioni, grafica e layout editing) e di incrementare la velocità di preparazione di layout sofisticati e la qualità di presentazione. In ambito *marketing* si ricorre alla promozione attraverso pagine web, alla vendita elettronica con form e pagamenti in rete.

Lo "**scenario 2**" vede la nascita dell'**editoria elettronica primitiva** (dal 1990). Nuove tecnologie di stampa permettono *printing on demand* nel senso che gli editori possono pubblicare piccoli quantitativi e rispondere velocemente a richieste impreviste, offrendo un servizio immediato al consumatore (trasferimento dei costi della produzione a stampa dal venditore all'acquirente). Non appena sono disponibili nuove informazioni gli autori e gli editori possono produrre edizioni rivisitate. Diventa possibile anche elaborare una qualche forma di *customisation* del prodotto per riflettere i bisogni dell'acquirente (includere l'identità dell'acquirente nel prodotto oppure fornire testi personalizzati con capitoli selezionati e appendici in sequenza). Data l'ampia disponibilità di schermi digitali e potendo distribuire via rete (pubblica e privata), emerge anche l'opportunità di incorporare informazione non stampabile nei documenti (suono, video, animazione). Inoltre l'affermazione del *desktop publishing* rende automatici (informatici) alcuni aspetti del processo di produzione dei materiali stampati. Molto maggiore è la possibilità di pubblicare materiale su mezzi diversi dalla carta (cd-rom e web).



Lo "**scenario 3**" si occupa dell'**editoria elettronica integrata** (dal 1998) ove la produzione *just in time* farà emergere i cosiddetti *living document* (suscettibili di continui miglioramenti). Gli autori attraggono visite ripetute ai contenuti mediante link ipertestuali nelle pubblicazioni (navigazione in *metadocumenti* e *documenti*). Esisterà la capacità di costruire *parti del documento* (tabelle e grafici) e anche *porzioni di testo*

mediante informazioni provenienti da interrogazioni verso basi di dati, di redigere indici e operare ricerche (basate su stringhe o su tecniche sintattiche). A tale scopo risultano indispensabili strumenti sofisticati rivolti alla semantica (thesauri generici, disciplinari e personali, analisi testuale e preferenze dell'utente). L'autore di contenuti dispone di software specializzato per preparare pubblicazioni digitali, dalla memorizzazione in unico formato master (XML, PDF) al recapito ai consumatori in formati multipli (HTML, Word, PDF, Postscript, XML).

Lo "**scenario 4**" contemplerà l'**editoria elettronica interattiva e adattativa** (dal 2000) che offrirà la possibilità di interazione tra la pubblicazione e l'utente (che potrà eseguire nuovamente una animazione, variare parametri e osservare l'impatto su una tabella o su un grafico). Gli effetti interattivi saranno ottenibili mediante software incorporato nei documenti (cd-i, televisione interattiva, applet). Non si può non pensare a software statistico in grado di elaborare dati contenuti in tabelle (XLS), per poi generare anche grafici.
Per quanto riguarda l'interazione tra utente e autore esistono già strumenti consolidati (posta elettronica, FAQ). Da sviluppare ancora, rimangono le possibilità di costruire documenti adattativi ricorrendo a formati particolari che soddisfano le esigenze del consumatore, tabulazioni e grafica, colori, font, dimensione e stile di stampa. Inoltre bisognerà permettere una personalizzazione basata su parametri comunicati dall'utente, parametri modificati dinamicamente da agenti software in base al comportamento dell'utente stesso.

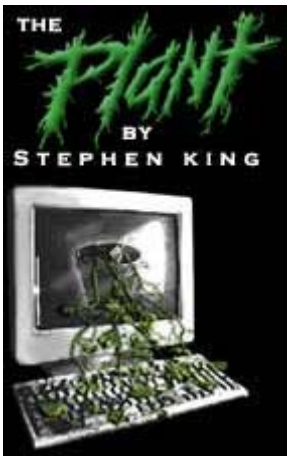
Editoria elettronica e commercio elettronico: un approccio digitale

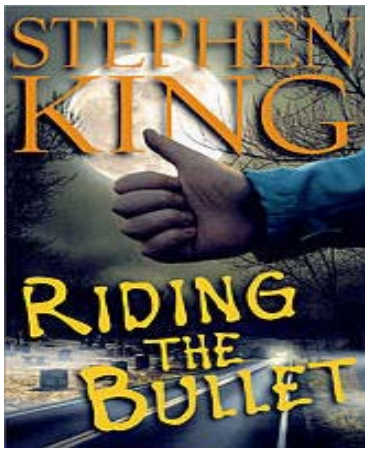
La *confluenza di web ed editoria* trae le sue origini dalla disponibilità di nuove tecnologie di stampa, dalla diffusione degli schermi multimediali e dalla possibilità di poter distribuire i contenuti via rete. Gli editori possono pubblicare piccoli quantitativi e rispondere velocemente a richieste impreviste, trasferendo nel contempo i costi della produzione a stampa dal venditore all'acquirente. Diventa possibile anche elaborare una qualche forma di *customisation* del prodotto per riflettere i bisogni dell'acquirente. Si concretizza anche l'opportunità di incorporare informazione non stampabile nei documenti (suono, video, animazione); diventa pratica comune anche pubblicare su mezzi diversi dalla carta (cd-rom e web).

La produzione *just in time* farà emergere i cosiddetti *living document* (passibili di continui miglioramenti) connessi strettamente alla capacità di costruire *parti del documento* (tabelle e grafici) e anche *porzioni di testo* utilizzando interrogazioni verso basi di dati, di costruire indici e fare ricerche.
L'autore dei contenuti dispone di software specializzato per preparare pubblicazioni digitali, dalla memorizzazione in unico formato master (XML, PDF) al recapito ai consumatori in formati multipli (HTML, Word, PDF, Postscript, XML).
Come si inserisce il "*libro virtuale*", o "***eBook***", nel contesto evolutivo sin qui tracciato? Perché gli eBooks potrebbero essere così interessanti dal punto di vista economico? Esistono diverse ragioni:

- **velocità di consegna del prodotto** (Internet);
- **portabilità** (anche i libri voluminosi diventano di dimensioni ridotte, se trasformati in eBook);
- **edizioni personalizzate**;
- **interattività e capacità di ricerca** (è molto semplice cercare un documento, inserire un segnalibro e aggiungere riferimenti incrociati e ipertestuali all'interno di eBook);
- **bassi costi per tutti** (gli editori possono ridurre i prezzi perché vengono a cadere i costi di stampa);

e tutte queste motivazioni si sono già manifestate in modi e forme diverse.
Il curioso fenomeno noto in rete come "*effetto Bullet*", è legato al successo enorme di un eBook recentemente pubblicato da Stephen King ("**Riding the Bullet**") e scaricato da più di 500.000 consumatori (picchi di 2.5 richieste al secondo).
Tale opera è stata commercializzata soltanto in formato digitale (eBook) e a costi molto limitati ($2.50US). La dimensione del libro è di 67 pagine (0.3MB) e il formato è PDF.
Tale autore si può considerare un precursore nell'uso dei nuovi formati, avendo già sperimentato differenti modalità di distribuzione e pagamento offerte dal web. Riguardo al suo ultimo lavoro, "**The Plant**", un romanzo a puntate pubblicato sul proprio sito, ha promesso di pubblicare le prime due parti: nel caso in cui la risposta comportamentale degli utenti fosse stata quella attesa (percentuale di paganti pari almeno al 75%), avrebbe pubblicato anche la terza parte. Il costo di ogni parte del romanzo scaricata è molto limitato (**1$US**) ma si richiede "*moralmente*" di non stampare copie extra da vendere o distribuire, contravvenendo alla legge sul copyright.





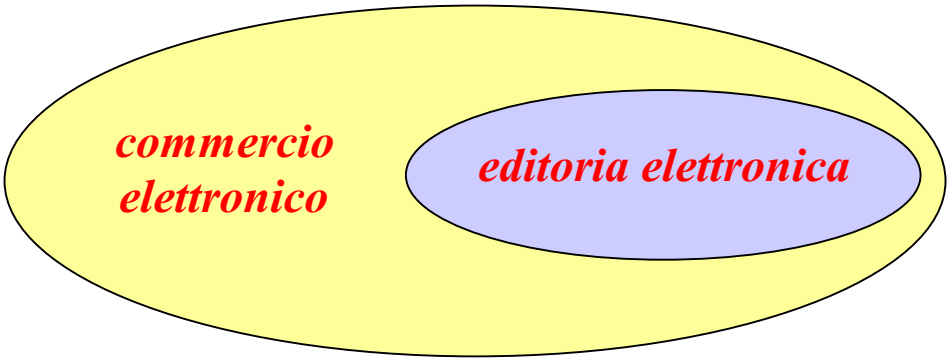
All'atto pratico ci sono stati 152132 download e hanno pagato 116200 utenti (76.38%). 93200 utenti hanno pagato subito e 23000 utenti hanno accettato di pagare dopo il download mantenendo l'impegno.

Tuttavia, l'esperimento di Stephen King ha evidenziato, oltre alla forza dell'editoria elettronica e del marketing editoriale via Internet) alcuni consistenti limiti. Uno di questi è stato accentuato da pirati informatici che hanno forzato i codici crittografici e reso il testo copiabile gratuitamente. Vengono sollevate quindi questioni fondamentali quali la tutela dei diritti d'autore e il mantenimento del margine per gli editori.
Il Sole 24 ORE inizierà la vendita online di libri elettronici, avviando la distribuzione in formato PDF di alcuni periodici (Guida Normativa, Guida al Diritto e Guida agli Enti Locali) direttamente sul sito web. Queste pubblicazioni sono oggi distribuite in abbonamento a oltre 95000 lettori ai quali, nel 2000, sarà offerta la disponibilità dell'edizione elettronica. Inoltre, in occasione della Fiera del Libro di Torino, il Sole 24 ORE ha presentato una nuova dimensione editoriale basata sul libro elettronico che si può scaricare dagli scaffali di una Libreria Virtuale (sito web).
Altre analoghe esperienze si possono reperire presso la **Libreria Editrice Vaticana** [39] e presso la casa editrice **Adelphi** che si accinge ad avviare la distribuzione e la vendita sul web di libri in formato elettronico in assoluta sicurezza e nel rispetto del diritto d'autore, proponendo una collana di brevi testi di narrativa e saggistica in formato digitale (PDF). La visualizzazione e la stampa dei documenti sarà possibile esclusivamente con la chiave di licenza venduta al singolo acquirente.
Un'altra esperienza significativa è legata a "**eDay**", un quotidiano online personalizzabile senza prezzo, non in vendita e non distribuito in edicola, che si riceve gratuitamente tramite Internet. Per leggere eDay basta essere in possesso di Acrobat Reader, software gratuito, che consente di visualizzare e stampare tale quotidiano digitale con la qualità di un normale quotidiano.
E' il primo quotidiano a essere progettato appositamente per integrarsi con Internet, anche se può essere stampato e letto come un giornale tradizionale. Inoltre è possibile usufruire dei seguenti servizi messi a disposizione dei lettori:

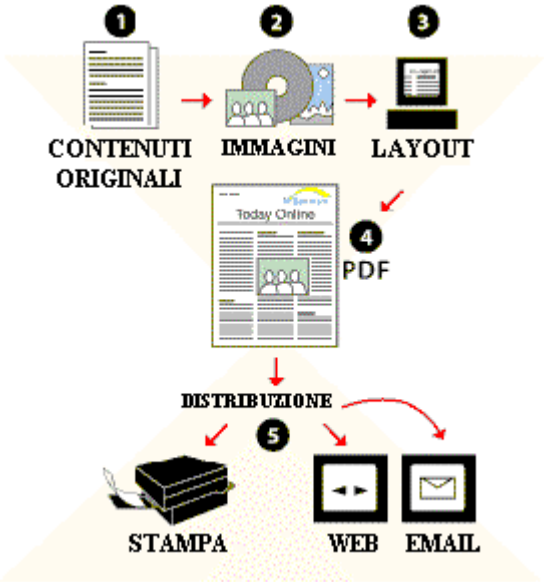
- personalizzare il giornale con il proprio nome e con le notizie desiderate;
- decidere quali pagine mettere nel giornale, in modo da avere un quotidiano veramente su misura;
- ricevere l'edizione personalizzata direttamente nella casella di posta elettronica.

Si tratta di una grande innovazione editoriale perché trasforma completamente il modello di business tradizionale del quotidiano, saltando costi, tempi e modalità legate alla carta, alla stampa e alla distribuzione.
L'idea del quotidiano personalizzato è applicabile pienamente anche al concetto di libro digitale. **McGraw-Hill**, una delle maggiori case editrici scolastiche, in autunno, dal suo sito **Primis Online**, offrirà titoli in formato PDF, personalizzabili e componibili.
Una funzionalità di "Add" permette di aggiungere ogni "*item*" desiderato nel libro personale. Il link "Your Complimentary Custom Book" invece permette di controllare la sequenza di item correntemente selezionata.
Infine, dopo aver scelto gli argomenti, bisogna stabilire la copertina e il titolo come pure le modalità di recapito. Si parla di "*shopping cart format*" e, non appena sono stati scelti i contenuti, il libro può essere stampato e pronto per la distribuzione in 6 settimane. Il settore del "*custom publishing*" è cresciuto del 20% annuo, rispetto al 5% dell'editoria tradizionale.
La **Random House**, la più grande casa editrice americana, ha lanciato una nuova struttura editoriale, **Atrandom**, che userà per le sue pubblicazioni solo strumenti digitali (libri elettronici) e stamperà su carta solo dietro richiesta dei clienti.
Il libro virtuale, o "*eBook*", si può collocare nell'ambito dei prodotti e dei servizi digitali ovvero prodotti e servizi che possono non solo essere acquistati (negoziati e pagati) utilizzando una *information infrastructure*, ma anche la relativa consegna può avvenire mediante gli stessi canali. I prodotti digitali esistono nel mondo digitale e potrebbero non aver bisogno di manifestarsi come oggetti fisici (immagini, film, animazione, audio, testo e software).

Esiste una naturale separazione tra *commercio elettronico* (include la consegna digitale del prodotto digitale) e altre forme di commercio condotte su Internet che implicano l'acquisto digitale di prodotti fisici (spediti per posta o consegnati mediante corriere). In un tale ambito è possibile vedere l'*editoria elettronica* come un sotto insieme del *commercio elettronico* (si parla di *digital publishing*).



La consegna digitale potrebbe aver luogo quando i prodotti vengono acquistati oppure quando viene scambiata informazione (ad esempio la compilazione di un questionario per la determinazione del *profilo acquirente*). Molto materiale digitale scambiato su Internet, infatti, non implica una transazione finanziaria.
La tendenza generale rilevabile negli standard editoriali innovativi è quella di prevedere comunque la compatibilità verso meccanismi di distribuzione di pubblicazioni digitali sia a pagamento che gratuiti. Una diffusione generalizzata sul *sito Web* istituzionale può essere rivolta all'utente che fornisce informazioni utili al monitoraggio utilizzando licenze d'uso gratuite. Nei casi in cui vi è un costo per le pubblicazioni digitali, la visualizzazione delle stesse è resa possibile soltanto mediante il possesso di una licenza d'uso acquistabile (non gratuita) da parte degli utenti.
In ogni caso, un tale contesto deve essere supportato da idonei processi produttivi editoriali. Un possibile processo editoriale (sistema di pubblicazione elettronico), ad esempio, potrebbe creare file in formato PDF e quindi permetterebbe a una "*singola sorgente*" di contenuti di essere pubblicata sul web e su cd-rom come anche su carta.



Nell'ambito di un tale processo editoriale, testo, immagini e collegamenti ipertestuali vengono combinati in layout modulari (passi 1-3), trasformati in PDF (passo 4) e pubblicati simultaneamente in diversi formati (passo 5).

D.R.M. (Digital Rights Management)

Il passaggio dal formato cartaceo a quello digitale permetterà a rivenditori e distributori una riduzione consistente dei costi legati a carta, inchiostro e stampa (costi collocabili nella tradizionale catena di produzione editoriale). Il *gap temporale* che intercorre tra la creazione intellettuale di contenuti e la produzione in formato digitale degli stessi, disponibili alla vendita, verrà drasticamente ridotto. Siccome il posizionamento dei prodotti offerti riguarda un *mercato virtuale* (*non fisico*), ne segue un miglioramento del *time-to-market*. Anche la capacità di recapito verrà ottimizzata perché i prodotti digitali sono tali che il loro acquisto può essere negoziato ed effettuato su Internet, ma anche il relativo recapito può avvenire tramite lo stesso mezzo. In breve tempo, molto materiale esistente verrà convertito in formato digitale, trasformando prodotti costituiti da atomi in prodotti costituiti da bit.
Questa trasformazione è applicabile anche alla *proprietà intellettuale* (*copyright*)?
Molte speranze concrete vengono riposte nello sviluppo di effettivi *meccanismi di pagamento*, di *modalità per esprimere diritti* su un materiale soggetto a proprietà intellettuale (*copyright object*) e di *protezione* di tale materiale in conformità ai diritti che sullo stesso vengono espressi.
Più in generale si parla di *gestione dei diritti digitali* ma anche di *schemi commerciali di copyright* per sottolineare la necessità di protezione delle informazioni digitali. A grandi linee, un *proprietario* (*owner*) deve poter assegnare o trasferire un *insieme di diritti* (*rights*) espressi su un oggetto digitale ad un'altra persona, sia essa editore, distributore, rivenditore o direttamente acquirente. Tale insieme di diritti viene trasferito mediante una *licenza* e permette di *effettuare delle operazioni* sull'oggetto in questione.
Le tecnologie D.R.M. (*Digital Rights Management*) permettono ai distributori di prodotti digitali di controllare l'accesso alla visualizzazione di tali prodotti mediante una sorta di *codifica personalizzata*. Agli utenti finali che hanno acquistato i diritti vengono fornite "*chiavi personali*" per la visualizzazione (o l'ascolto) dei prodotti, generalmente con una limitazione sulle funzionalità di copia, stampa e ridistribuzione. Tuttavia, per diffondere i contenuti ad una platea più ampia possibile, si può prevedere che il contenuto venga ampiamente distribuito – rendendolo disponibile in maniera generalizzata (*sito web*) – ma possa essere visionato soltanto mediante una licenza d'uso valida (*voucher*), venduta separatamente ad ogni acquirente ed eventualmente connessa all'ambiente informatico del medesimo. L'editore può generare chiavi che specificano i *permessi* da attribuire al file utente (*abilitazione alla stampa*, *annotazioni*, *editing*, *selezione di testo e grafica*) che tengano però conto anche delle *caratteristiche dell'ambiente* nel quale il documento può essere utilizzato. In questo modo è possibile adattare i termini o condizioni di vendita a esigenze commerciali individuali (*cosa* è necessario per accedere al contenuto, *dove* dirigere l'utente per acquistarlo, *come* accedere al contenuto) e dar vita a politiche commerciali diversificate.
Il rivenditore dovrebbe considerare le varie modalità con le quali consegnare i *diritti di accesso*. Un sistema di sottoscrizione (*abbonamento*) potrebbe permettere di scaricare (*download*) presso un sito web direttamente i contenuti, magari in corrispondenza di precisi eventi temporali, oppure inviare gli stessi contenuti tramite posta elettronica. Una pianificazione editoriale diffusa a priori potrebbe rendere i contenuti accessibili soltanto per un certo periodo di tempo.
Una volta che all'utente viene assegnato il diritto d'accesso, i distributori devono assicurarsi anche che tutti gli attori del processo di creazione, produzione e distribuzione vengano pagati equamente per l'uso del prodotto. Per quanto riguarda le modalità di compenso si adotta un *pagamento a priori* del copyright per quantità fisse, per licenze che abilitano alla riproduzione di un certo numero di copie e comunque in

tutti i casi in cui il mercato di prodotto è stabile. Un *pagamento a posteriori* viene scelto quando la domanda del prodotto non è prevedibile.

Tecnologie e strumenti D.R.M.

Negli *schemi commerciali di copyright* un *proprietario* (*owner*) deve poter assegnare o trasferire un *insieme di diritti* espressi su un oggetto ad un'altra persona. Tale insieme viene trasferito mediante una *licenza* e permette di *effettuare delle operazioni* sull'oggetto in questione. Possibili diritti possono essere di **trasporto** (copia e trasferimento), di **rendering** (display, stampa e riproduzione), di **derivazione** (estrazione, inclusione, modifica, adattamento) e di **backup** (protezione contro danni al prodotto e/o alla licenza).
Un programma server che utilizza la sintassi dell'eXtensible Markup Language (XML) può essere utilizzato appunto per assegnare i diritti di accesso e generare le chiavi crittografiche per i file PDF. Ad un utente generico poi si permette di scaricare dal web file cifrati per poi sbloccarli e visualizzarli sul PC.
I rivenditori possono decidere i requisiti per l'accesso ai contenuti e dove gli acquirenti possono acquistare tali diritti. Allo scopo si utilizza un numero di identificazione legato al computer dell'utente (disco rigido), il nome utente, ID della CPU oppure la data corrente. L'utente stesso può decidere quali requisiti condividere per effettuare l'acquisto.

```
<?xml version="1.0"?>
<…>
<content>
<seller name="…" id="…" url="http://…"/>
<document-set url="http://…/ebusiness.fdf"
              title="E-Business: Roadmap for Success">
<document id="fb00025"/>
</document-set>
<buyer name=""/>
<transaction id="200002151628426697" utc="20000215163502"/>
<license>
<eq name="CPU" value="CPU-PC0573A77735B02AAAZ71"
    key="JuSHj5Yy5Izo60zC4solG6NKJy3iwmx7Xcytr86n
         +hp8H/UGyMIyvIMwTMRaJIwgDbWHhvgciYsWf3Iu
         lp4kIAcWjV7exhHwK7/yNO4UNn2gRC+uiJI/cFkO
         IgR7cDvLDU3yQV+fKRY5EiIy65peeJTTFFrPBKCtnVS1gYLGOqk="/>
</license>
</content>
<signature value="Y9ySBTwlAW6pDcg4Ijvq32L7f7y6sBjLmcpw3Wx7bYaCxFx
                  CZiu9Hm5VKtsRFyflU5/TRa9xNBjG1qhjUI7V2VH7CPEbH4
                  VDV7z5MKzPsoP12k9JIp+XRYERtXXhk7ISRy8frNjzp+ZDW
                  V0EsQaiC8qsDrRCWuZ3tTaTeuE0YKc="
          cert="…"/>
</…>
```

Il processo DRM prevede varie fasi e la più significativa è quella nota come "*locking del documento*". Le variabili di input richieste per il *processo di locking* sono il file PDF, URL del venditore, ID del venditore, ID del documento e i permessi di accesso al file. Gli output del processo di locking sono il file PDF bloccato (cifratura a 56 bit) e una *chiave di titolo* ovvero un dizionario cui viene apposta la firma digitale (1024 bit). Viene generata una chiave che identifica univocamente il contenuto ed è poi richiesta per sbloccare il documento. Tali chiavi devono essere disponibili presso il sito web del venditore (URL del venditore), nel senso che – siccome nel file PDF vengono codificati il *nominativo del venditore* e l'indirizzo del proprio sito web – utenti sprovvisti di

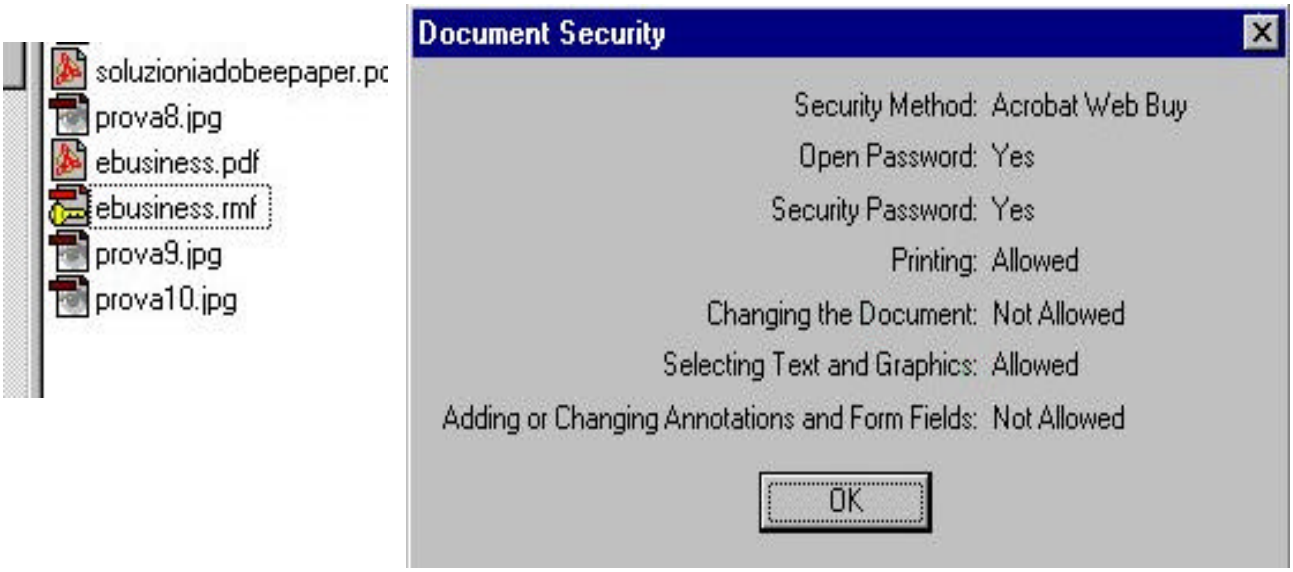
licenza d'uso vengono indirizzati verso il sito del venditore per acquistare i necessari diritti di accesso.
Per ogni acquirente viene creata una licenza d'uso contenente la chiave di titolo del documento e tale licenza viene legata all'ambiente di elaborazione dell'acquirente. La licenza d'uso (voucher) deve essere memorizzata nella stessa directory del documento. E' un file firmato digitalmente (*XML 1.0* e *certificati X.509*) che contiene un ID del documento, i permessi e una *espressione di ambiente*.
Il venditore determina quale insieme delle *variabili di ambiente* relative all'elaboratore del consumatore utilizzare per ottenere un adeguato *livello di sicurezza* (**CPU ID** unico per ogni computer, **HD ID** unico per ogni disk drive presente nel computer, **network disk ID** poiché ogni disk drive condiviso in rete è unico, **removable disk ID** unico per floppy disk, CD-ROM, Iomega Zip e Jaz, **tempo** come *Grrenwich Mean Time* utile per controlli temporali e scadenze di licenze d'uso, **username** quale classica procedura di login sul computer). L'acquirente può inviare tutte o soltanto alcune di tali variabili e, di conseguenza, il venditore decide se le variabili inviate sono sufficienti per bloccare il file PDF oppure declina la richiesta di acquisto.
Sviluppi futuri permetteranno di includere *smart card*, *indirizzi IP*, *impronte digitali*, *crittografia a chiave pubblica* e *firma digitale*.
Delle *espressioni AND/OR arbitrarie* composte di vari identificatori precedentemente proposti possono descrivere le circostanze utili affinché una particolare licenza possa aprire un documento bloccato:

- **ALLOW if USER="Carchesio" AND TIME ≥"1-Sep-2000:00:00"**
- **ALLOW if CPU-ID="PC0573A77735B02AAAZ71_0" OR CPU-ID="PC0813A07795E02AAAK95_0"**
- **ALLOW if CPU-ID="PC0573A77735B02AAAZ71_0" AND HD-ID="PP2221A2DB60(MS_DOS6)"**
- **ALLOW if RD-ID="Click!" AND TIME ≤"16-Sep-2000:00:00" AND ORDER_TIME ≥"1-Sep-2000:00:00"**

ad esempio, l'ultima espressione proposta permette il recapito della versione digitale di un libro la cui versione a stampa è già stata ordinata (consegna prevista per il 16 Settembre 2000).
Tali espressioni forniscono un elevato grado di flessibilità e di conseguenza permettono modelli di business sia tradizionali che creativi. In ogni caso viene consigliata la creazione di *espressioni con almeno due alternative* perché, se l'acquirente rompe l'hardware associato ad una alternativa (**CPU-ID="CPU-PC0573A77735B02AAAZ71_0"**), può ancora utilizzare la seconda alternativa per aprire il documento (**HD-ID="HD-PP2221A2DB60(MS_DOS6)"**). Ovviamente, quando si predispongono alternative bisogna fare uso di "**OR**". Una tale scelta permetterebbe all'acquirente di aggiornare la propria CPU oppure il proprio HD senza dover comprare di nuovo una apposita licenza d'uso.
Nel documento bloccato, come pure nella licenza d'uso associata, vengono codificati anche i *permessi di accesso*. Attualmente tali permessi sono quelli di **print** (possibilità di stampare pagine), **change** (possibilità di apportare modifiche al documento), **select** (possibilità di selezionare testo o grafica su una pagina per copiare tali informazioni su un'altra applicazione) e **annotation** (possibilità di usare tool per annotazioni e per markup). I permessi sono applicabili soltanto sull'intero documento.
Una caratteristica relativa ai permessi deve essere abilitata sia nel documento PDF che nella licenza d'uso. Alcuni editori usano la tattica di abilitare ogni cosa nel file PDF e disabilitare delle funzionalità nella licenza d'uso. Qualora un consumatore volesse acquisire maggiori permessi, dovrebbe comprare una licenza d'uso dotata di maggiori abilitazioni. In ogni caso non occorre cifrare di nuovo il file PDF.
Se viene richiesta la funzione di "**Open**", senza disporre della necessaria licenza d'uso, viene di fatto impedita l'apertura del file PDF bloccato. Ovviamente l'utente viene

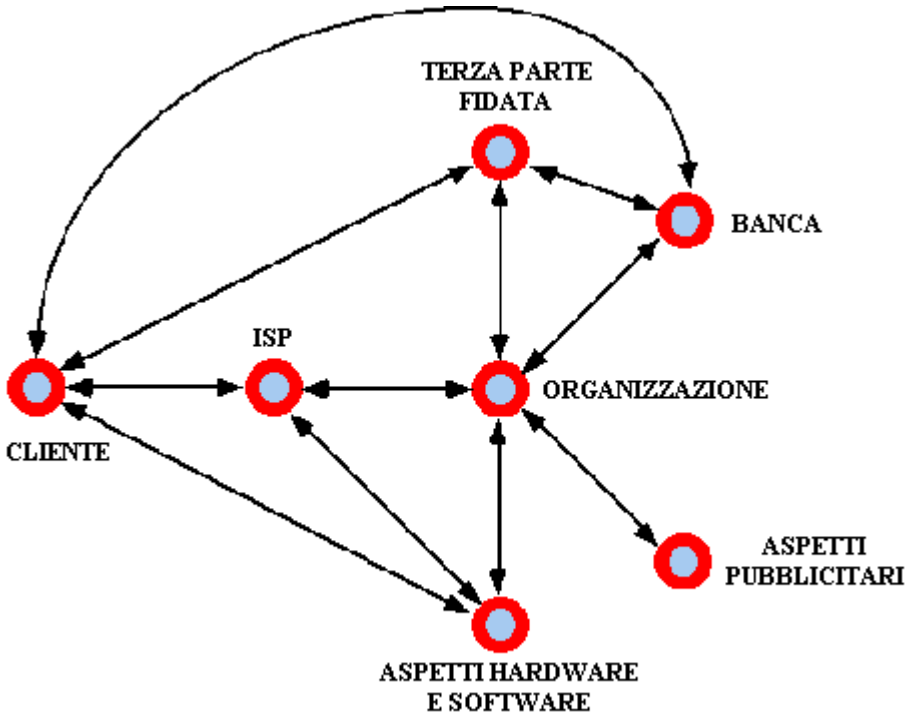
avvisato della mancata apertura (*assenza di una licenza d'uso*). Tuttavia viene offerta la possibilità di reperire tale *licenza d'uso* mediante Internet.
L'utilizzo della funzione "**Go On-line...**" provoca la navigazione sul *sito del venditore*, dove viene richiesto all'utente di fornire un insieme di informazioni necessarie ad ottenere la licenza d'uso. Nel caso specifico le informazioni richieste saranno *codificate* assieme alle caratteristiche del documento da visionare, precedentemente scaricato, in modo da impedire che lo stesso documento possa essere distribuito a terzi illegalmente.

Si è visto che occorre fornire *informazioni sensibili* (codice identificativo CPU e codice identificativo HD, indirizzo di posta elettronica, nominativo), anche se viene data assicurazione sul trattamento confidenziale delle stesse informazioni.
A questo punto all'utente viene consegnata della *licenza d'uso*. Un nuovo utilizzo della funzione "**Open**" effettivamente rende il file PDF *disponibile in lettura*. Se si osserva con attenzione la directory ove è presente il documento da visionare (**ebusiness.pdf**), si nota la presenza di un nuovo file (**ebusiness.rmf**). Tale file è proprio la licenza d'uso (*vocheur*), anche se in gergo tecnico si parla di "***Rights Managment File*** (***.rmf**)". La licenza viene gestita dal reader Acrobat effettuando una traduzione di tipo XML (*parsing*).



Distribuzione di prodotti editoriali Istat in formato digitale via web

Un modello generale di riferimento per la costituzione di una infrastruttura dedita al *commercio online di prodotti digitali* (e conseguente *distribuzione*) prevede:



tuttavia, per la costituzione di una infrastruttura dedita alla sola "*distribuzione dei prodotti editoriali Istat in formato digitale via web*", si può non tener conto di:

- **terze parti fidate** (rapporti che Istat dovrebbe intrattenere con una o più *Certification Authority* nell'ambito di problematiche legate all'utilizzo della *firma digitale* conforme alle regole tecniche AIPA);
- **banche** (rapporti che Istat dovrebbe intrattenere con una o più istituzioni bancarie nell'ambito di problematiche legate alla gestione di *transazioni economiche* per il *pagamento* dei prodotti editoriali venduti via web);
- **pubblicità** (distribuzione effettuata senza intermediari).

Una infrastruttura semplificata, rispetto al modello generale, dovrebbe prevedere:

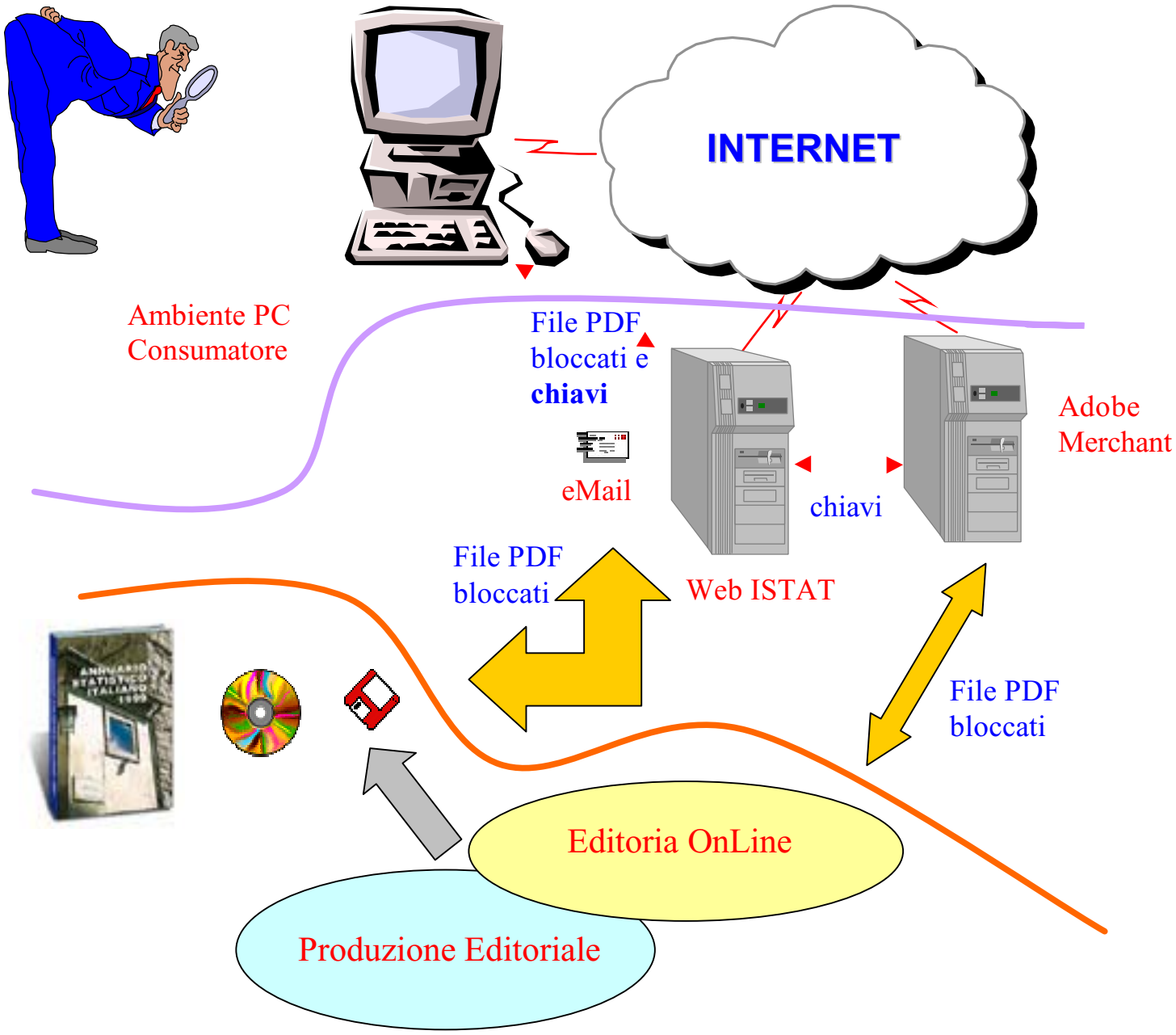
- **ambito del consumatore**;
- **ambito della distribuzione via web**;
- **ambito dei servizi editoriali**.

Per quanto riguarda l'ambito del consumatore, bisogna tener conto della *gestione degli abbonamenti*, dell'*ambiente informatico* (posta elettronica, connessione Internet presso un ISP, browser), del rilascio gratuito di un *reader* adeguato (Acrobat Reader), della realizzazione di una *indagine conoscitiva* mirante a identificare la propensione all'acquisto di prodotti editoriali online (eventualmente da condurre sugli abbonati attuali oppure tramite questionario libero sul sito Istat).

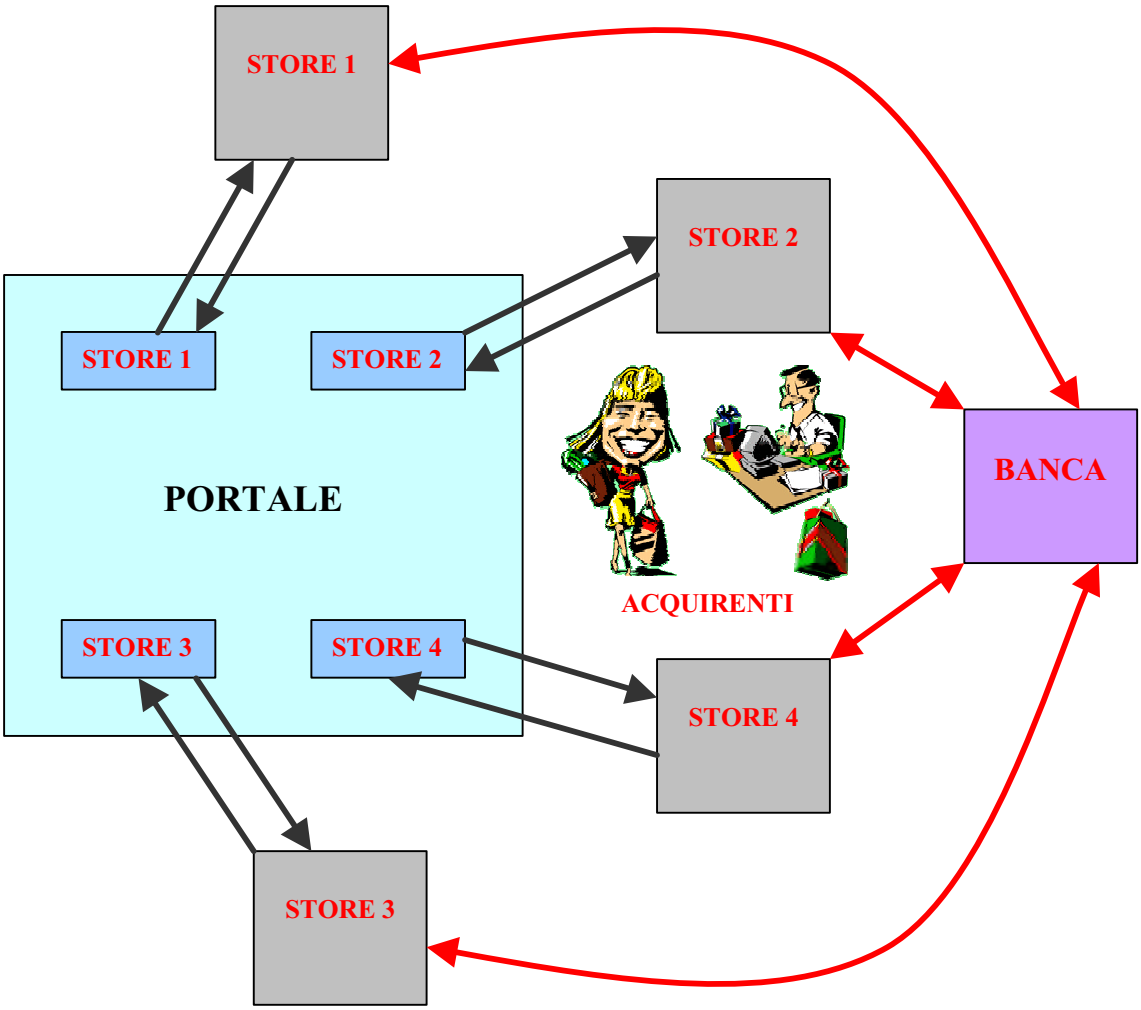
La distribuzione dei prodotti via web dovrebbe tener conto della *gestione dei diritti digitali* stabilendo opportuni meccanismi (informazioni quali CPU-ID e HD-ID sembrano troppo invasive dal punto di vista della sicurezza del PC del consumatore), di un database degli abbonati per fornire informazioni utili per la realizzazione di licenze d'uso, della *fornitura di servizi a supporto della vendita* (*printing on demand* delle pubblicazioni voluminose e *preview* nel periodo di tempo che intercorre tra l'acquisto del volume cartaceo e la relativa consegna al comsumatore, *mastering on demand* di CD cumulativi annuali, *try-and-buy* di contenuti parziali gratuiti).

Sono i servizi editoriali (*produzione editoriale*, *editoria online*) che, dopo aver condotto una analisi di ampio respiro per individuare i *prodotti editoriali classici* idonei alla *conversione nel formato digitale* (Annuario, Bollettino Mensile di Statistica, Istat Informa, Tabelle di Consultazione), ne predispongono la *distribuzione*. Tale analisi

dovrebbe determinare le effettive possibilità di convertire in PDF gli output editoriali attualmente prodotti (tramite *scanning* e *OCR* se trattasi di materiale cartaceo oppure tramite *import* da formati già digitali).



Qualora la distribuzione dei prodotti editoriali classici avvenisse mediante l'utilizzo di un *intermediario*, bisognerebbe verificare la disponibilità presso tale intermediario di strutture e tecnologie web idonee (*portal*) alla distribuzione online di prodotti digitali.
Anche in questo caso si potrebbe prescindere dalla presenza di un interlocutore bancario oppure – volendo offrire un sistema di pagamento evoluto (carta di credito) - la gestione delle transazioni finanziarie potrebbe essere affidata all'intermediario stesso mediante forme di contrattazione apposite. Rispetto allo schema proposto, vi sarebbe una "*sezione Istat*" sul *portale* dell'intermediario di distribuzione (ad esempio "*store2*") che rimanda direttamente al "*sito Istat*" dove effettivamente sono presenti le pubblicazioni digitali da distribuire.



Sondaggi e propensioni di acquisto di prodotti digitali

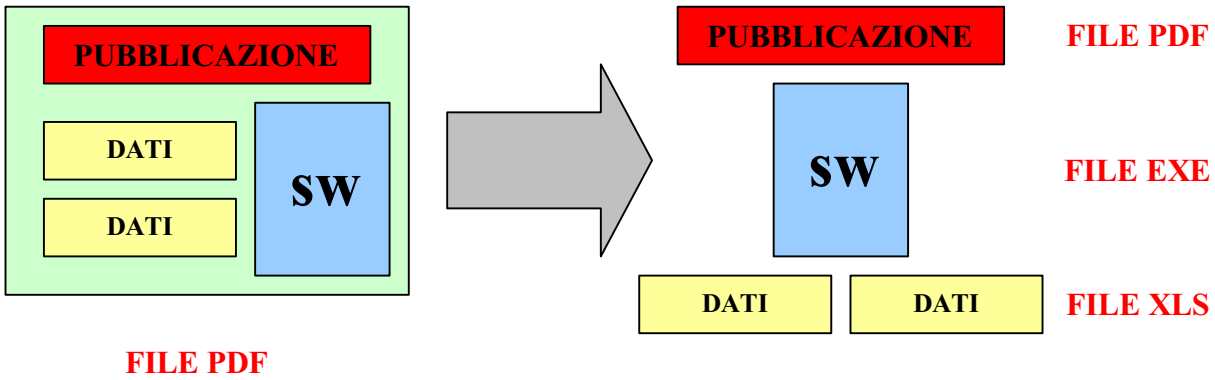
Il carattere *fortemente innovativo* di una eventuale **distribuzione di prodotti editoriali Istat in formato digitale via web**, collocabile nel più ampio panorama costituito dal *commercio elettronico*, suggerisce una attenta valutazione del *contesto nazionale* in termini di:

- diffusione di infrastrutture telematiche (possibilità di navigare in Internet);
- propensione all'acquisto di prodotti e servizi digitali;
- grado di fiducia riposto nell'acquisto e nel relativo pagamento online;
- valutazione dei costi ritenuti accettabili rispetto alla categoria di prodotto e/o servizio.

Delle utili indicazioni possono essere mutuate da precedenti sondaggi, anche se l'Istat provvederà - a breve termine – alla rilevazione conoscitiva presso la propria utenza in relazione alle tematiche sopra indicate.
Ad esempio, rispetto alle varie tipologie possibili di *offerta di prodotti in Internet*, classificata per settori, si può rilevare un consistente gradimento per le voci inerenti ai servizi (14.1%), all'editoria (13.2%) e alla pubblica amministrazione (11.8%).
Il commercio elettronico viene considerato con molto favore (36.4%) o comunque ad un livello di interesse abbastanza significativo (47.3%).
In relazione alle attuali tecnologie di sicurezza, invece, un acquisto tramite Internet viene considerato generalmente poco sicuro (50.9%), anche se un certo grado di fiducia è abbastanza presente (32.7%) e, in pochi casi molto sentito (3.6%).

Vantaggi e svantaggi del formato PDF

A latere delle interessanti possibilità offerte dall'utilizzo del *formato PDF* bisogna opportunamente valutarne vantaggi e svantaggi. Tra gli *svantaggi* bisogna collocare sicuramente la caratteristica che vede PDF essere un *formato proprietario*, con tutte le conseguenze del caso (acquisto di software e di licenze adeguate, formazione del personale, aggiornamento costante delle nuove versioni).
Tra i dubbi, un quesito importante riguarda, poi, la *effettiva diffusione del formato PDF*. A tale riguardo si può notare che, sul sito Istat, attualmente circa il **25%** dei link si riferisce a documenti PDF (**HTML**, **WORD**, **EXCEL**). Presso il **sito EUROSTAT** le pubblicazioni gratuite sono in formato PDF mentre quelle a pagamento sono disponibili sia in formato HTML che PDF. In generale, non limitandosi all'ambito statistico nazionale ed europeo, è possibile rilevare una consistente diffusione di materiale in formato PDF. Questo aspetto implicitamente equivale a dire che esiste una capacità generalizzata di produrre in formato PDF, anche se tale affermazione dovrebbe essere rapportata alla *effettiva realtà editoriale di ogni singola organizzazione*.
Una *prospettiva futura* incoraggiante viene fornita dalla Deliberazione AIPA n° 24/98, relativa alle "*Regole tecniche per l'uso di supporti ottici*" nell'ambito della archiviazione di documenti nella Pubblica Amministrazione. Tale delibera individua, per i "*documenti formati all'origine su supporto informatico*", la possibilità di essere trasferiti sul supporto ottico di memorizzazione, senza passaggio su supporto cartaceo, in un formato conforme allo *standard SGML* (HTML) oppure ai formati PDF, AFP e Metacode.
E' evidente che per la Pubblica Amministrazione ci si attende, per i prossimi anni, una evoluzione "*dal cartaceo al digitale*" basata sull'utilizzo di una serie di formati tra cui PDF. Di conseguenza nel breve periodo è lecito attendersi un notevole aumento di materiale in formato digitale e, in proporzione, un aumento altrettanto sensibile di materiale digitale in formato PDF.
Tra i vantaggi, da un lato si rileva che la documentazione in formato PDF, rispetto alla corrispondente documentazione in altro formato, presenta una *riduzione di dimensione*. Nel caso Istat tale riduzione risulta variabile (44% - 81%). Dall'altro è impossibile non considerare che il software che permette la visualizzazione di file PDF (Acrobat Reader) è disponibile *gratuitamente* sul web ed è stato - ad oggi - distribuito in 180 milioni di copie.
Infine, indiscusso vantaggio per la peculiarità delle pubblicazioni Istat, il formato PDF permette l'*inclusione di formati nativi* (.XLS) ma anche l'*estrazione di formati nativi*.



Ciò permette, ad esempio, di allegare ad una pubblicazione PDF eventuali dati statistici (tabella EXCEL) e un eventuale software di elaborazione degli stessi dati (un file eseguibile .EXE). Il prodotto finale consiste in un unico file PDF e come tale può essere distribuito per mezzo di Internet. Il destinatario di un tale file PDF può, in fase di consultazione, estrarre sia i dati che il software per elaborare tali dati in modalità

nativa ovvero memorizzando nel proprio ambiente di elaborazione (PC) non uno ma più file distinti.

XrML (eXtensible rights Markup Language)

Attualmente esistono due formati per eBook (PDF e una variante di HTML e XML), ognuno supportato da un diverso *eBook reading device*. Tentativi di unificazione sono reperibili in un recente rilascio di "*Open eBook 1.0 Publication Structure Specification*", specifiche di pubblico dominio per un formato comune di eBook basato su XML e HTML. Tuttavia queste specifiche presentano ancora dei problemi da risolvere, ad esempio la *mancanza di sicurezza* insita nell'assenza di meccanismi rivolti alla protezione del copyright e alla possibilità di effettuare transazioni sicure.
La chiave di volta per la crescita è costituita dalla disponibilità di meccanismi "*Digital Rights Management*", basati sulla semplice idea di vendere e/o comprare opere digitali mediante sistemi sicuri per mezzo dei quali i diritti possono essere inglobati (attached) nelle stesse opere. Si avverte l'esigenza di un *linguaggio standard* per proteggere diritti e imporre termini e condizioni in modo che le opere digitali possano essere attribuite ai relativi autori, vendute e successivamente fruite su varie piattaforme. Le regole interpretate dai computer devono essere totalmente obiettive e consistenti per imporre un *contratto digitale*. Quando viene richiesta una operazione non prevista dalla licenza, un sistema sicuro rifiuta di eseguire tale operazione.
Tale standard è ora disponibile come *XrML* (eXtensible Rights Markup Language). XrML supporta l'editoria e la vendita di materiale digitale come pure l'accesso e il controllo di utilizzo per uno scambio non gratuito di contenuti digitali sicuri. XrML semplifica l'uso tramite regole basate su un linguaggio e una sintassi consistenti, è licenziato gratuitamente, dispone di una architettura aperta, offre scalabilità e personalizzazione, si può estendere e ha la capacità di integrarsi sia con sistemi esistenti che con sistemi da sviluppare. XrML supporta molti standard importanti tra i quali corre l'obbligo citare SSL e crittografia a chiave pubblica.
Il modello XrML è applicabile ad arbitrari contenuti digitali, siano essi una singola opera, un'opera vista come le singole parti componenti oppure una collezione di opere. Un qualsiasi contenuto digitale viene identificato utilizzando un oggetto. Vari diritti di utilizzo vengono poi specificati sul medesimo contenuto. Il codice identificativo del contenuto e i diritti specificati, assieme alle relative meta informazioni (chi ha generato il documento XrML, quando e per chi), vengono firmati digitalmente per assicurare la loro integrità. I documenti XrML vengono rappresentati da elementi XrML con la seguente struttura:

```
<XrML>
      <BODY>
            (ISSUED)?
            (TIME)?
            (DESCRIPTOR)?
            (ISSUER)?
            (ISSUEDPRINCIPALS)?
            (WORK)?
            (AUTHENTICATEDDATA)?
      </BODY>
      (SIGNATURE)?
</XrML>
```

dove "?" denota zero o una occorrenza, "+" denota una o più occorrenze e "*" denota zero o più occorrenze.

All'interno dell'elemento radice XrML, è presente un elemento obbligatorio (mandatory) BODY e un elemento opzionale SIGNATURE (la firma digitale usata per assicurare l'integrità). L'elemento BODY consiste di una descrizione opzionale di un'opera digitale e di alcune opzionali meta informazioni sul documento XrML. L'elemento TIME è un intervallo temporale nel quale il documento XrML è valido. L'elemento ISSUED individua il momento in cui tale documento è stato creato. L'elemento DESCRIPTOR è una descrizione del documento stesso.
L'elemento ISSUER indica chi ha originato il documento XrML, mentre ISSUEDPRINCIPALS è una lista di destinatari cui lo stesso documento è rivolto. L'elemento WORK definisce un'opera digitale e i relativi diritti. Infine, l'elemento AUTHENTICATEDDATA raccoglie informazioni da fornire ad un'applicazione che deve elaborare tale documento XrML.
Questa struttura ad alto livello fornisce flessibilità nel definire associazioni tra ISSUER, ISSUEDPRINCIPALS e WORK. Se tutti gli elementi citati sono presenti, allora il documento XrML definisce una associazione del tipo "ISSUER rilascia il documento XrML ai ISSUEDPRICIPALS specificando diritti di utilizzo su WORK".
XrML specifica un'opera digitale e i relativi diritti di utilizzo mediante una struttura quale:

```
<WORK>
        (OBJECT)
        (DESCRIPTION)?
        (CREATOR)*
        (OWNER)?
        (DIGEST)*
        (PARTS)?
        (CONTENTS)?
        (COPIES)?
        (COMMENT)?
        (SKU)?
        (RIGHTSGROUP | REFERENCEDRIGHTSGROUP )+
</WORK>
```

dove l'elemento OBJECT è l'oggetto che può essere usato per identificare l'opera mentre DESCRIPTION ne è una descrizione. L'elemento CREATOR individua l'autore dell'opera e l'elemento OWNER è il proprietario della stessa (potrebbero coincidere).
L'elemento DIGEST è una impronta dell'opera usata per assicurarne l'integrità. Specifica un algoritmo di hash (ALGORITHM), i relativi parametri (PARAMETER) e un valore di impronta opzionale (VALUE).

```
<!ELEMENT DIGEST (ALGORITHM,PARAMETER*,VALUE?)>
```

L'elemento SIGNATURE contiene le informazioni sulla firma digitale apposta sul BODY di un documento XrML, ovvero elementi obbligatori DIGEST e VALUE ma anche elementi opzionali come ALGORITHM e PARAMETER. L'elemento DIGEST fornisce informazioni hash del BODY mentre ALGORITHM specifica l'algoritmo usato per generare e verificare la firma digitale. L'elemento VALUE è il valore attuale della firma digitale, codificato in accordo alle specifiche del relativo attributo. Base64 è un metodo di codifica comunemente utilizzato. Quando ALGORITHM è assente, l'algoritmo associato con le chiavi pubbliche e/o private dell'elemento ISSUER nel BODY viene selezionato.

```
<!ELEMENT SIGNATURE (DIGEST, ALGORITHM?, PARAMETER*, VALUE)>
```

Un esempio potrebbe essere:

```
<SIGNATURE>
<DIGEST>
<ALGORITHM>SHA1</ALGORITHM>
<PARAMETER name="codingtype">
<VALUE encoding="string">surface-coding</VALUE>
</PARAMETER>
<VALUE encoding="base64" size="160">OtSrhCKWHI … </VALUE>
</DIGEST>
<VALUE encoding="base64" size="512">A7qsNTFT2 … </VALUE>
</SIGNATURE>
```

Con PARTS si indica una lista di tutte le parti della medesima opera che potrebbero avere diversi diritti di utilizzo, costi e condizioni. Ogni parte è a sua volta un'opera.
L'elemento CONTENTS indica dove il contenuto corrente è posizionato rispetto all'opera globale, utile quando il contenuto coperto da diritti è incorporato in un'opera digitale. L'elemento COPIES specifica il numero di copie dell'opera mentre COMMENT contiene un qualsiasi commento sull'opera o sui relativi diritti di utilizzo.
L'esempio che segue riguarda un "eBook" che non ha costi associati per la visualizzazione (*playing*), ma che ha 5$ di costo per poter creare copie digitali. Sono stati specificati anche due diritti di "Print", entrambi legati ad una stampante sicura. Il primo diritto di Print può essere esercitato se l'utente possiede un particolare ticket già pagato, mentre il secondo diritto di Print richiede un costo fisso di 10$. L'esempio assume che il libro digitale può essere trasmesso ad un PC utente esercitando il diritto di Copy prima, e che poi l'utente possa visionare o stampare (play or print) l'opera in base alle proprie esigenze esercitando i relativi diritti.

```
<?XML version="1.0" encoding="UTF-8"?>
<!DOCTYPE Work SYSTEM "xrml.dtd">
<XrML>
        <WORK>
        …
        <BUNDLE>
                <ACCESS>
                <SECURITYLEVEL name="Physical" value="4"/>
                <SECURITYLEVEL name="Kernel" value="5"/>
                </ACCESS>
                <FEE><MONETARY>
                <ACCOUNT>
                <ACCOUNTTO id=" Account-Jones-Pub248afdoiu4398"/>
                </ACCOUNT>
                </MONETARY></FEE>
        </BUNDLE>
        <RIGHTSLIST>
        <COPY>
        <FEE><MONETARY><PERUSE value="5"/></MONETARY></FEE>
        </COPY>
        <TREANSFER></TREANSFER>
        <DELETE></DELETE>
        <PLAY></PLAY>
        <PRINT>
        <PRINTER><CERTIFICATE>
        <AUTHORITY id="Digital Property Trust"></AUTHORITY>
        <PROPERTYPAIR name="Class" value="Trusted Printer-6"/>
        </CERTIFICATE></PRINTER>
```

```
        <FEE><TICKET type="Prepaid-Print">
        <AUTHORITY id="Jones Publishers"></AUTHORITY>
        </TICKET></FEE>
        </PRINT>
        <PRINT>
        <PRINTER><CERTIFICATE>
        <AUTHORITY id="Digital Property Trust"></AUTHORITY>
        <PROPERTYPAIR name="Class" value="Trusted Printer-6"/>
        </CERTIFICATE></PRINTER>
        <FEE><MONETARY><PERUSE value="10"/></MONETARY></FEE>
        </PRINT>
        </RIGHTSLIST>
        </RIGHTSGROUP>
        </WORK>
</XrML>
```

Conclusioni

Il "*libro virtuale*" ("*eBook*") si inserisce a pieno titolo in un contesto evolutivo del processo di produzione editoriale e concorre fortemente all'attuazione del principio "*electronic dissemination first*". Standard editoriali innovativi sono sempre più caratterizzati dalla compatibilità verso meccanismi di distribuzione di pubblicazioni digitali che consentono sia la modalità a pagamento che quella gratuita.
Nell'ottica dei continui miglioramenti da apportare al sistema integrato di diffusione, e quindi ai processi di produzione editoriale, è opportuno prevedere un monitoraggio continuo del bacino di utenza (questionari online, gestione di abilitazioni e password per gli accessi a banche dati) i cui esiti poi vadano ad alimentare idonei sistemi informativi di marketing.
Di conseguenza, per diffondere le pubblicazioni Istat ad una platea che sia la più ampia possibile, una diffusione generalizzata e gratuita sul *sito Web* istituzionale (*luogo privilegiato di diffusione*) può avere luogo attuando un processo di mutua collaborazione che vede l'utente fornire informazioni utili al monitoraggio e l'Istat mettere a disposizione licenze d'uso gratuite per la fruizione di pubblicazioni in formato digitale.
Tuttavia, nei casi in cui è necessario proteggere la *proprietà intellettuale* (*copyright*) oppure si ritiene opportuno attribuire un costo alle pubblicazioni digitali, la visualizzazione delle stesse è resa possibile soltanto mediante il possesso di una licenza d'uso acquistabile (non gratuita) da parte degli utenti.
L'evoluzione del processo editoriale quindi colloca l'*editoria elettronica* in una posizione strettamente connessa con le attività di *commercio elettronico* (*digital publishing*). Diventa possibile adattare termini e condizioni di distribuzione (gratuita) e/o vendita (costo) a esigenze individuali (*cosa* è necessario per accedere al contenuto, *dove* dirigere l'utente per acquistarlo, *come* accedere al contenuto) e dar vita a politiche commerciali diversificate.
Le tecnologie D.R.M. (*Digital Rights Management*) permettono una efficace gestione delle licenze d'uso mediante una *codifica personalizzata* basata sulla *crittografia*, ormai assunta a strumento effettivo, economico e versatile. Agli utenti finali che hanno acquistato i diritti vengono fornite "*chiavi personali*" per la visualizzazione dei prodotti, generalmente specificando i *permessi* da attribuire al file utente (*abilitazione alla stampa*, *annotazioni*, *editing*, *selezione di testo e grafica, copia, ridistribuzione*).
Linee evolutive interessanti sono attese dal linguaggio *XrML*. Da un lato, essendo *XrML* parte di XML, risulta possibile reperire e integrare oggetti eterogenei provenienti da sistemi informativi diversi. I cosiddetti *living document* saranno connessi strettamente alla capacità di costruire *parti della pubblicazione* (tabelle, grafici, *porzioni di testo*)

utilizzando i risultati di interrogazioni verso basi di dati. Dall'altro, essendo incapsulati in *XrML* meccanismi rivolti alla protezione del copyright e alla possibilità di effettuare transazioni sicure (SSL e crittografia a chiave pubblica), diventa semplice imporre regole (*contratto digitale*) sull'utilizzo di tali oggetti eterogenei.

BIBLIOGRAFIA

1. M. McLuhan, "La galassia Gutemberg: Nascita dell'uomo tipografico", Armando, Roma 1976

2. R.L. Rivest, A. Shamir, L. Adleman, "A Method for Obtaining Digital Signatures and Public-Key Cryptosystems", Communications of the ACM 21,2 (1978), 120--126

3. Comunicazione e società oggi e domani – "Il Rapporto Mc Bride sui problemi della comunicazione nel mondo", ERI-Rai, Roma 1981

4. D. Scavetta, "La metamorfosi della scrittura: dal testo all'ipertesto", La Nuova Italia Scientifica, Firenze 1992

5. S. Brand, Media Lab. "Il futuro della comunicazione", Baskerville, Bologna 1993

6. D. Charney, "The Impact of Hypertext on Process of Reading and Writing", Modern language association, New York 1993

7. G.P. Landow, "Ipertesto. Il futuro della scrittura", a cura di Bruno Bassi, Baskerville, Bologna 1993

8. M. Bangemann, "Europe and the global information society (The Bangemann Report)", Recommendations to the European Council, 26 May 1994, Brussels (Belgio) – Information Society Project Office (http://www.ispo.cec.be/infosoc/backg/bangeman.html)

9. P. Ghislandi, "Oltre il multimedia", Franco Angeli, Milano 1995

10. R. Ridi, "Internet in biblioteca", Bibliografica, Milano 1996

11. M. Bryan, "Sgml and Html Explained", Addison-Wesley, New York 1997

12. F. Cairncross, "The Death of Distance: How the Communications Revolution Will Change Our Lives", Harvard Business School Press, Boston 1997

13. ITU-T X-Series Recommendation, Data Networks and Open System Communications, Directory, Authentication Framework, X.509, Giugno 1997

14. D.P.R. 10 novembre 1997 n. 513 - Regolamento recante criteri e modalità per la formazione, l'archiviazione e la trasmissione di documenti con strumenti informatici e telematici, a norma dell'articolo 15, comma 2, della legge 15 marzo 1997, n. 59

15. R. Clarke, "Electronic Publishing: A Specialised Form of Electronic Commerce", 10th International Electronic Commerce Conference, Bled, Slovenia, June 1997

16. E. Guidotti, "Internet e comunicazione", Franco Angeli, Milano 1997

17. A. Lioy, F. Maino, M. Mezzalama: "Secure document management and distribution in an open network environment", ICICS'97 - International Conference on Information and Communications Security, Beijing (China), November 11-13, 1997

18. M. Morelli, "La comunicazione d'impresa e la promozione dell'immagine", Franco Angeli, Milano 1997

19. P. Atzeni, A.O. Mendelzon, G. Mecca, "The World Wide Web and Databases", International Workshop WebDB'98, Valencia (Spain), March 27-28, 1998

20. P. Atzeni, "Basi di dati e nuove tecnologie di rete a supporto della produzione statistica", Quarta Conferenza Nazionale di Statistica, 12 novembre 1998, Roma

21. G. Di Carlo, "Internet marketing. Strategie di mercato e di comunicazione sul web", Etas Libri, Milano 1998

22. C. Bendin, M. Ceresa Gastaldo, "Impresa e Comunicazione. Dalla privacy al web", Erga, Genova 1999

23. A. Cadioli, "Dall'editoria moderna all'editoria multimediale. Il testo, l'edizione, la lettura dal Settecento ad oggi", Edizioni Unicopli, Milano 1999

24. F. Carlini, "Stile del web. Parole e immagini nella comunicazione di rete", Einaudi, Torino 1999

25. P. Ferri, "La rivoluzione digitale. Comunità, individuo e testo nell'era di Internet", Mimesis, Milano 1999

26. R. Clarke, "Key Issues in Electronic Commerce and Electronic Publishing", Information Online and On Disc 99, Sydney, 19 - 21 January 1999

27. V. Lo Moro, "Innovative ways of meeting user need on official statistics", Issues in Data Dissemination, Sixth CEIES Seminar, Manchester, 28-29 January 1999

28. G. Mottura, "Pubblica amministrazione e comunicazione. L'Istat e le nuove frontiere della comunicazione online", Università di Roma "Tor Vergata", 1999

29. N. Negroponte, "Essere digitali", Saggi Paperback, Sperling & Kupfer Editori, Febbraio 1999, Roma

30. "An Introduction to Web Farming", Febbraio 1999 (http://webfarming.com/intro/default.html)

31. W. Diffie, "Public Key Infrastructure: Today and Tomorrow", Workshop su "La firma digitale – un fatto .COMpiuto", 29 Marzo 1999, Roma

32. R. Clarke, "Electronic Trading in Copyright Objects and Its Implications for Universities", Australian EDUCAUSE'99 Conference, Sydney, 18-21 April 1999

33. G. Carchesio, "Dal documento cartaceo al documento elettronico", collaborazione editoriale per il sito "Jure Page – Diritto&Internet", Giugno 1999

34. "Adobe and Digital Content for eCommerce", Adobe Systems Incorporated, September 1999

35. "Public Key Criptography Standards" (PKCS#1 - PKCS#6 - PKCS#7 - PKCS#10), RSA Laboratories (http:/www.rsasecurity.com/rsalabs/pkcs/)

36. J. Rifkin, "L'era dell'accesso: la rivoluzione della new economy", Mondadori Editore, Marzo 2000

37. V. Lo Moro, "Diffusione dell'informazione statistica nell'era di Internet", Giornale del Sistan n° 14/2000

38. "Don't Let Your Files Leave Home Without Em" (http://www.digitalpayloads.com/)

39. "Providing assurance for e-Business digital content using PKI based digital signatures", Assured WebSite (http://www.elock.com/products/WebSite/)

40. "The Official Stephen King Web Presence" (http://www.stephenking.com/)

41. Libreria Editrice Vaticana (http://www.vaticanbookstore.com/)

42. XrML (eXtensible rights Markup Language) (http://www.xrml.org/)

43. "Build a book online" (http://www.mhhe.com/primis/)

WORKSHOP

Editoria elettronica per la diffusione dell'informazione statistica

La realizzazione di "e-book" con l'utilizzo di acrobat

A. Braghin e P. Crivelli

LA REALIZZAZIONE DI "e-book" CON L'UTILIZZO DI ACROBAT

A. Braghin[1] e P. Crivelli[2]

Una questione molto dibattuta negli ultimi anni per chi si occupa in vario modo di editoria riguarda il modo e il mezzo con cui il prodotto editoriale viene predisposto e poi rilasciato e se e con quale velocità le pubblicazioni cartacee saranno sostituite da modalità di divulgazione elettroniche e tramite Internet.

Nell'arco di un decennio, ed in stretta connessione con l'evoluzione della tecnologia, a fianco delle pubblicazioni cartacee, a volte in alternativa ad esse, si sono sviluppate le modalità di diffusione su floppy disk (più spartane e più compatte) dapprima, poi su CD-Rom (supporti molto più capienti dei floppy e corredate in genere di applicazioni per la ricerca e la navigazione nell'insieme dei dati ivi contenuti) e, parallelamente, si è sviluppata la modalità di diffusione *online* attraverso la rete Internet.

Ciascuno dei suddetti mezzi di comunicazione ha le sue specificità: le pubblicazioni su carta hanno il più basso livello di accesso nel senso che un volume è qualcosa di fisicamente tangibile e di immediata consultazione anche per chi non ha familiarità con la tecnologia informatica. Una pubblicazione su supporto magnetico (ad esempio un CD-Rom) necessita invece per essere consultata di un elemento intermedio, un computer, e di una conoscenza tecnica specifica da parte dell'utilizzatore. In compenso però un CD-Rom, che può contenere anche migliaia di pagine, si può mettere comodamente in tasca, non ingombra e non pesa; inoltre può contenere percorsi di consultazione guidati o agevolati, animazioni, link ipertestuali ad altre informazioni correlate temporalmente o comparate geograficamente. Entrambi, sia il volume cartaceo che il supporto magnetico sono accessibili solo a chi li possiede fisicamente.

Con la diffusione *online* le informazioni consultabili in rete sono disponibili per tutti quelli che si connettono alla rete (e che non possiedono fisicamente l'oggetto) ovviamente se gratuite. In questo caso la conservazione del documento dipende da chi gestisce il sito ove questo è presente.

Per un ente come l'Istat, nel definire una politica di diffusione, non è facile optare per una modalità o l'altra dando luogo a conseguenti rilevanti modifiche dei processi produttivi.

Una buona soluzione in questo senso consiste nel poter evitare di fare scelte drastiche e poter valutare e decidere a seconda delle necessità. Ciò vuol dire concretamente potersi porre, come produttori dei dati, nella condizione di realizzare la stessa pubblicazione su carta, su CD-Rom e su Internet con il minimo lavoro aggiuntivo a partire da un unico input fornito da chi produce i dati.

[1] Tecnologo, presso l'Istat - Servizio Sviluppo di prodotti per la diffusione - Stampa modelli ed editoria censuaria.
[2] Primo Tecnologo presso l'Istat - Servizio Sviluppo di prodotti per la diffusione - Editoria elettronica e centro stampa.

In questa direzione, l'utilizzo delle potenzialità del prodotto Acrobat può determinare una rilevante innovazione di processo ad anche una innovazione di prodotto e contribuire sensibilmente a realizzare l'indicazione strategica " electronic first" che l'Istituto Nazionale di Statistica si è data.
Il livello di accesso all'utilizzo del prodotto, al "come fare", è molto basso e quindi può essere esteso con relativa facilità anche in un ambiente di lavoro in cui sono presenti figure professionali molto diverse fra loro.

Che cos'è Acrobat/PDF

Acrobat è un prodotto software per l'editoria elettronica che raggruppa un insieme di moduli dedicati a specifiche funzioni. E' stato sviluppato dalla società Adobe, proprietaria anche di altri software di ampio utilizzo come Photoshop, Illustrator, Page Maker, dei formati di file PDF (Portable Document Format) e POSTSCRIPT, e di un ampio insieme di font tipografiche.
Acrobat include diversi componenti, come riportato in tabella 1, secondo le finalità d'utilizzo:

Tabella 1

Finalità	Componente	Descrizione
Visualizzare	- Acrobat Reader - Acrobat Player	Consultazione a video e interazione con documenti PDF
Produrre	- Acrobat Exchange - Acrobat Capture and import Plug-ins - Acrobat Distiller - PDFWriter	Conversione elettronica di file o documenti cartacei in formato PDF; inserimento di link ipertestuali
Indicizzare	- Acrobat Catalog	Generazione di indici per ricerche semplici e complesse

Acrobat produce degli output in formato PDF. Tale formato costituisce uno standard (suffisso ".PDF") che la Adobe ha a suo tempo registrato e reso di pubblico dominio le specifiche di sviluppo. A partire da software come Word, Excel, Access, e più specifici come Quark XPress e tutti i prodotti della Adobe tra i quali Page Maker, Photoshop, Illustrator, è possibile salvare i documenti in formato ".PDF". C'è infine una interazione stretta con il formato di stampa PostScript.
Con riferimento alla tabella 1 il modulo ***Reader***, scaricabile gratuitamente dalla rete Internet o spesso contenuto nei CD-Rom con cui vengono diffusi documenti PDF, consente di visualizzare, stampare, copiare, aggiungere note e appunti personali al documento, se consentito dall'autore.

Il modulo ***Exchange*** oltre alle funzioni del Reader consente di aggiungere oggetti, di effettuare modifiche testuali e di formattazione del documento, aggiungere link ipertestuali, proteggere il documento con password, inibendo all'utilizzatore le funzioni di cui sopra.

Il ***Distiller*** interpreta e converte file PostScript e EPS (Encapsulated Post Script) effettuando una compressione che, rispetto al PostScript, può essere anche 10 o 20 volte inferiore.

Le opzioni di processo di Distiller controllano la compressione, l'incorporazione dei font, la gestione dei colori e molte altre proprietà del file PDF creati con Distiller. Acrobat prevede tre gruppi di opzioni predefiniti che possono essere utilizzati nelle situazioni più comuni, ossia:

a) Le opzioni *Schermo Ottimizzato* riguardano i file PDF che verranno visualizzati sul WorldWide Web o su una rete Intranet o che verranno distribuiti con un sistema di posta elettronica per la consultazione su schermo. Questo gruppo di opzioni garantisce la compatibilità con Acrobat versione 3.0 e successive;
b) Le opzioni *Stampa Ottimizzata* riguardano i file PDF che verranno inviati a una stampante o a una copiatrice digitale, che verranno pubblicati su CD o che verranno presentati a un cliente come bozza di stampa. Questo gruppo di opzioni garantisce la compatibilità con Acrobat versione 4.0 e successive;
c) Le opzioni *Prestampa Ottimizzata* riguardano i file PDF per cui verrà eseguita una stampa finale di alta qualità. Questo gruppo di opzioni garantisce la compatibilità con Acrobat versione 4.0 e successive.

Non è possibile modificare questi tre gruppi di opzioni predefiniti ma è possibile utilizzarli come punto di partenza per creare dei gruppi personalizzati o definire e salvare dei gruppi di opzioni *ex novo*. Il ***Distiller*** va utilizzato per tutti quei documenti che contengono informazioni sulla gestione dei colori o che utilizzano colori indipendenti dal dispositivo. Inoltre, è necessario conservare le caratteristiche PostScript se i file devono essere inviati a un centro di prestampa o ad un laboratorio professionale per la stampa di alta qualità, oppure se si desidera utilizzare un processo *batch* per convertire i file PostScript in PDF o se si vogliono incorporare font asiatici nel file stesso.

Il ***PDFWriter*** consente di stampare rapidamente i documenti in formato PDF. Per creare un file PDF con *PDFWriter* è possibile utilizzare il trascinamento della selezione o il comando stampa della selezione con cui è stato creato il documento. Esso è indicato per i documenti aziendali semplici che contengono prevalentemente testo. Se i documenti da convertire sono ricchi di grafica è preferibile utilizzare *Distiller*.

Il modulo ***Catalog*** consente di generare un insieme di indici, salvati in un file con lo stesso nome e suffisso ".PDX", sia di singoli documenti che di insiemi di documenti PDF. Consente inoltre di effettuare ricerche testuali libere, ricerche booleane, fonetiche e di prossimità.

Il *Catalog* consente di creare indici full-text di documenti o raccolte di documenti PDF. Dopo aver creato un indice è possibile usare il comando *Ricerca* per cercare rapidamente nell'intera libreria l'informazione di cui si ha bisogno. Un indice *full-text* è un data base che contiene tutto il testo di un documento o di un gruppo di documenti. La creazione di un indice *full-text* è possibile solo con file di tipo *PDF Normale e PDF Immagine* originale con testo nascosto, cioè di documenti digitalizzati a cui Acrobat applica il riconoscimento ottico dei caratteri (OCR).

Nella *nota tecnica **1*** è riportata la procedura per l'installazione del prodotto.

Il modo più semplice per trasformare un documento in versione Acrobat PDF consiste nel dare un semplice comando di stampa del documento selezionando opportunamente una stampante virtuale, Acrobat per l'appunto, in modo che venga prodotto un file in un modo più o meno simile alla elaborazione delle pagine che una stampante laser effettua.
Per poter fare ciò occorre disporre non soltanto della componente *Reader* del prodotto, che consente di "leggere" oggetti realizzati in Acrobat da altri, ma anche del pacchetto *Exchange* che consente di aggiungere e togliere pagine al documento Acrobat iniziale, definire dei link a pagine specifiche, aggiungere icone, oggetti, immagini, note personali di commento e così via in modo da fornire al documento quel valore aggiunto che la versione cartacea non ha.
I software originari con cui si definisce l'input del documento da trasformare successivamente in ".PDF" possono essere molteplici e riguardare:

- programmi di video scrittura *Word*; *Word Perfect*, *Wordstar*, eccetera;
- fogli elettronici *Excel*. *Lotus*;
- programmi adatti all'impaginazione professionale come *QuarkXpress*, *Page Maker*; *Desktop Publishing*; *Corel Ventura;*
- editor di immagini e fotoritocco professionale come *Adobe PhotoShop*, *Illustrator,* eccetera;
- editor per il Web come *Front Page*, *Adobe Page Mill*.

Ad essi possono essere affiancati, dopo la trasformazione del documento in PDF, compilatori con cui confezionare il ".PDF" in maschere e pacchetti di installazione, scrivendo qualche riga di codice di programma.

L'innovazione di processo

L'Istat potrebbe trarre un notevole vantaggio dall'inserimento del prodotto Acrobat nei processi di produzione e di rilascio dell'informazione statistica al fine di migliorare la diffusione elettronica ed in particolare quella effettuata via Internet.
A completamento di quanto già detto riguardo al prodotto, occorre prendere atto che Acrobat è uno standard *de facto*; è il modo più utilizzato per diffondere elettronicamente testi voluminosi, manuali tecnici ed altra documentazione ad ampia diffusione. Inoltre, i prodotti software più diffusi per l'editoria elettronica e la prestampa, a cominciare dal mondo Microsoft Office, prevedono una uscita standard nel formato PDF ottenibile in modo estremamente semplice.
A partire da ciò si evidenziano alcuni concetti che, se sviluppati ed attuati, possono portare all'Istituto considerevoli vantaggi in termini di efficienza e di risultato:

- Con Acrobat è molto semplice realizzare la struttura di concatenazione tra gli elementi (pagine, oggetti) che compongono una pubblicazione dandole in tal modo il valore aggiunto della ipertestualità;
- Con Acrobat si può semplificare e velocizzare l'iter di approntamento e correzione delle bozze di stampa;

- La diffusione su Internet di documenti in formato PDF offre significativi vantaggi.

Attualmente in Istat le fasi di elaborazione ed aggregazione dei dati relativi alle indagini correnti sono risolte soprattutto con programmi SAS mentre per la predisposizione di file e tavole statistiche formattate, e quindi pronte per la diffusione, si utilizza prevalentemente Excel.
Testi in Word e tavole in Excel costituiscono la modalità di diffusione per le collane editoriali *Informazioni-dati congiunturali* (soltanto su floppy disk o CD-Rom), *Annuari* e *Informazioni*. Queste ultime due vengono stampate presso tipografie esterne fotoriproducendo gli originali forniti da Istat.
Per le pubblicazioni generali, le collane *Argomenti* ed *Essays* o per quelle in quadricromia come *Indicatori statistici* gli input Word ed Excel vengono importati in ambiente di impaginazione professionale Xpress. Per questi si attiva tra il servizio produttore e la tipografia che impagina i volumi e per il tramite della struttura Istat che gestisce le pubblicazioni, l'iter di correzione delle bozze. In entrambi i casi l'utilizzo di Acrobat potrebbe essere molto vantaggioso.
Nel primo processo produttivo la versione finale delle tavole da fotoriprodurre potrebbe essere confezionata in Acrobat ed inviata così alla tipografia, con maggiori garanzie rispetto al risultato voluto, dal momento che la formattazione Excel a volte può variare leggermente da computer a computer.
Nel caso di pubblicazioni soltanto su supporto magnetico (collana *Informazioni dati-congiunturali*) queste potrebbero essere arricchite con elementi di ipertestualità ottenendo con poco lavoro ulteriore dei prodotti editoriali migliori, più guidati e più gradevoli.
Nel secondo processo produttivo, che comporta di norma una impaginazione esterna all'Istat, l'iter di correzione potrebbe essere più veloce, più semplice ed efficace.

Strutture ipertestuali

Acrobat è un software per costruire una rete ad *albero* da una *radice* iniziale attraverso i *rami* verso le *foglie*, oppure una struttura reticolare che prevede rinvii più complessi da una pagina all'altra.
L'importante è disporre dei testi (o predisporre le immagini) di tutte le *foglie* o pagine finali e delle pagine intermedie, che sostanzialmente elencano le scelte di navigazione consentite. Il resto è un lavoro di "cucitura" dei singoli pezzi tramite la definizione di link e collegamenti a cui si provvede seguendo le opzioni del prodotto ma senza scrivere alcuna riga di codice di programmazione.
Quindi, nella ipotesi più semplice se si dispone di un documento preesistente e sequenziale questo lo si può trasformare in formato PDF. Seguendo il procedimento contrario, si può progettare un'applicazione per Acrobat prevedendo la concatenazione delle pagine fin dall'inizio e quindi predisponendole opportunamente.
I collegamenti e i link consentono di passare a posizioni specifiche che possono trovarsi nello stesso documento, in altri documenti elettronici o su siti Web. Grazie ai collegamenti e ai link, l'utente può accedere immediatamente alle informazioni correlate. Questi, inoltre, possono essere utilizzati

per eseguire azioni come la riproduzione di un clip audio o video, la lettura di un articolo o l'invio di un modulo.
Collegamenti e link vengono creati con lo strumento *Collegamento* e *Modulo* e possono essere sia visibili che invisibili. E' inoltre possibile collegare file di diversa estensione (Excel, Word, ecc.) al documento PDF. Questo consentirebbe di dotare le pubblicazioni ISTAT di documenti coadiuvanti e di fornire all'utenza finale strumenti di lavoro eterogenei ed esaustivi (tabelle in Excel, lettere e documenti in Word e quant'altro). Il requisito fondamentale, per la lettura di un file PDF composto da più file di altra estensione, è quello di avere a disposizione l'applicazione in grado di leggere i file collegati al file PDF "madre". In altre parole, per leggere un file ".XLS" collegato ad un documento PDF, l'utente dovrà avere istallato l'applicazione Excel, così dicasi per Word, ecc. Inoltre, il percorso dato al file dal "creatore" del file non dovrà mai essere modificato pena l'impossibilità di aprire il file collegato al PDF.
L'incorporazione in un file PDF di un clip audio e video, per esempio, può trasformare un semplice documento PDF in uno strumento multimediale avanzato. I filmati e i suoni possono essere riprodotti quando si attiva un collegamento, un segnalibro oppure un pulsante.
Un'altra opzione consente di assegnare i filmati e i suoni come azioni di pagina, che vengono attivate all'apertura o alla chiusura di una pagina, così come la possibilità di adottare il "pieno schermo" cioè senza vedere la barra dei comandi di Acrobat all'apertura del documento e/o di adattare il documento PDF alla finestra del computer in cui lo si sta visualizzando. In questo caso si dovrà necessariamente prevedere, per una migliore visualizzazione, che le dimensioni del documento siano correlate a quelle delle schermo in linea di massima.

La semplificazione dell'iter di correzione delle bozze di stampa

E' possibile, con l'utilizzo di Acrobat, semplificare ed ottimizzare la fase di prestampa e di revisione delle bozze.
Quando la fotocomposizione di volumi viene effetuata presso una tipografia esterna o un centro interno specializzato per tale funzione, di norma l'invio di originali e bozze avviene per mezzo di un corriere o utilizzando servizi postali. La correzione viene effettuata sul documento cartaceo.
Salvando il documento in formato PDF con Acrobat, la tipografia incaricata dell'impaginazione potrebbe inviare ai servizi committenti e/o all'ufficio preposto al controllo redazionale le prime bozze allegando il documento ad un messaggio di posta elettronica. Tale accorgimento porterebbe, in questa fase, ad una sicura ottimizzazione del processo di produzione in quanto:

a) Si eliminerebbero i tempi e i costi legati all'uso di un corriere incaricato di recapitare le bozze a stampa al servizio committente;
b) Si eviterebbero correzioni di bozza (d'autore e non) riportate a penna o pennarello sull'impaginato cartaceo. Infatti, le eventuali correzioni dovute ad aggiunte o ad anomalie sorte nella fase di impaginazione potrebbero essere corrette direttamente dal/dagli autori sul documento PDF attraverso le "note di testo".

Nella barra degli strumenti sono disponibili tre tipi di strumenti di annotazione: *strumenti Nota*, *strumenti Grafici* e *strumenti Testo*. Ognuno dispone di un'ulteriore barra visualizzabile facendo clic sull'icona. Tali strumenti consentono di associare commenti personalizzati ai documenti PDF in vari modi; tuttavia, ogni strumento offre un metodo di comunicazione delle informazioni unico.

Le note possono essere collocate ovunque nelle pagine dei documenti PDF. E' inoltre possibile adattare lo stile e il formato delle note al documento e al tipo di commento che si desidera associare.

c) Inoltre, l'utilizzo dello strumento "note" eviterebbe la digitazione, da parte dell'operatore, di interi blocchi di testo o parte di essi, abbassando notevolmente i margini di errore connessi, per ovvi motivi, con tale operazione. Anche in questo caso si verificherebbe un notevole abbassamento dei tempi e dei costi, in quanto il numero di bozze verrebbe drasticamente ridotto con l'abbassarsi delle possibilità di errore spesso imputabili all'operatore e al correttore (mancata chiarezza dei segni di correzione, fraintendimenti, scarsa leggibilità della grafia, imputazione errata durante la digitazione, ecc.). Attraverso l'invio delle bozze per posta elettronica, la struttura preposta avrebbe la possibilità di creare, nell'immediato, un archivio/database informatizzato da utilizzare per le operazioni di controllo e di informazione da inviare, in tempo reale, agli organi preposti interni all'Istituto;
d) Infine, giunti al "Visto si Stampi", l'opera sarebbe immediatamente disponibile per la pubblicazione in rete oppure off-line. Anche in questo caso si opererebbe un abbattimento dei tempi richiesti per la stampa in offset.

Una importante caratteristica del prodotto Acrobat è quella di effettuare **la separazione dei colori per la stampa in Offset,** necessaria per gestire la fase finale del processo di stampa. E' importante, a tale proposito, tenere presenti alcuni accorgimenti nella preparazione del file in PDF affinché rispondano a questo tipo di esigenza (***vedi nota tecnica 2***).

La diffusione su Internet di documenti in formato PDF

I file in formato PDF come pure in HTML hanno entrambi delle caratteristiche che li rendono molto efficienti nella diffusione in rete Internet. Tra queste la compatibilità multi piattaforma, l'interattività, l'ipertestualità, la multimedialità, la compattezza. Acrobat, che non è nato come prodotto integrato con i web browser come il linguaggio HTML, è tuttavia perfettamente compatibile con i maggiori browser come Internet Explorer e Netscape Navigator.

Per rendere un documento PDF più efficiente in rete sia per la visualizzazione che per il download, è opportuno "ottimizzarlo" con la funzione "Batch optimize" del modulo Acrobat Exchange. La funzioni di ottimizzazione ordina e identifica tutti gli elementi presenti nel documento: testi, immagini, funzioni interattive, ecc. e crea un indice in grado di puntare tutti gli elementi; inoltre compatta ulteriormente il documento, in modo che il Web server invii all'utente solo le pagine richieste invece che tutto il documento, che potrebbe essere anche molto corposo. L'ottimizzazione

consente ad Acrobat di gestire il documento in modalità *Progressive rendering* con cui le pagine vengono visualizzate o scaricate una dopo l'altra senza dover attendere il caricamento in memoria di tutto il documento. Con il *Progressive rendering,* o restituzione progressiva, vengono visualizzati prima i testi, poi i grafici, gli oggetti interattivi, quindi le immagini ed infine le font. Mentre l'utente sta visualizzando una pagina le pagine successive vengono scaricate e sono già disponibili anche se non tutto il documento è scaricato.
La modalità *page-on-demand* della versione 3.0 consente all'utente di richiedere la pagina che desidera e gestisce la memoria *cache* in modo da memorizzarvi tutte le pagine precedenti a quella desiderata e renderle poi immediatamente disponibili su richiesta o per il normale *scrolling* del documento.
La combinazione di *Progressive rendering* e *page-on-demand downloading* consente all'utente di visualizzare un file PDF di 100 Mbytes come uno di 100 pagine con pochissima differenza di prestazioni e tale da rendere più conveniente il formato PDF per file di grandi dimensioni rispetto all'analogo HTML.

Un prototipo sperimentale di e-book

Per la distribuzione su CD-Rom o su Floppy Disk è possibile creare delle applicazioni, ad esempio in Visual Basic, che consentano, non solo di consultare uno o più documenti in formato PDF tramite un'interfaccia amichevole e personalizzata, ma anche di distribuire file aventi un'altra estensione e permettere altresì l'installazione di altre applicazioni.
Nel caso specifico è stato predisposto un Form, installabile attraverso una procedura tipo, in Visual Basic in cui sono stati incorporati, come oggetti OLE, diversi file, il file PDF contenente il formato elettronico e "navigabile" di una pubblicazione già stampata dall'Istat in precedenza, due file di testo e l'applicazione di Acrobat Reader 4.0 (italiano) per Windows 95/98/NT che, come è noto, è indispensabile per consentire la lettura di file PDF, ovvero file generati dai programmi di impaginazione elettronica di Adobe, come Acrobat Exchange ed Acrobat Distiller.
La versione presentata nella casella "applicazione" del *Form* contenente l'icona rossa è la più compatta, priva del motore di ricerca.
Per installare il programma "ISTAT E-BOOK", concepito a scopo dimostrativo, è sufficiente cliccare due volte sull'apposita icona di SETUP. Questa avvia il programma di installazione ("setup") dell'applicazione. Essa può essere lanciata dal menu START (Windows 95/98) del PC, selezionando la cartella PROGRAMMI e successivamente la cartella, creata *ad hoc* in modo automatico durante l'installazione, ISTAT E-BOOK.
Dunque, lanciando l'applicazione compare un *Form* che permette di visualizzare le caselle contenenti i file, l'applicazione di Acrobat Reader, precedentemente illustrata, e un pulsante che permette l'uscita dal programma. Sulla sinistra del *Form* è stata posta l'immagine del volume contenuto nel file PDF, oggetto centrale dell'applicazione. Tutti i file contenuti nelle caselle e l'applicazione di Acrobat sono attivabili con un doppio clic.

Un programma così concepito permetterebbe altresì di inserire tanti *Form* quante sono le pubblicazioni elettroniche che si voglio divulgare. Creando, ad esempio per collane o per annate, serie complete di pubblicazioni elettroniche consultabili e "navigabili" da un solo CD-Rom.
La versione di ISTAT E-BOOK si disinstalla anch'essa con la procedura standard, ossia:
"Menu Avvio, Impostazioni, Pannello di controllo, Installazione Applicazioni". Nella lista che appare, si seleziona la voce "istat e-book", poi si "clicca" sul pulsante <Aggiungi/Rimuovi>. Questa operazione cancella tutti i file del programma, le icone nel menu Avvio e le modifiche apportate a Windows durante l'installazione.

Barra di navigazione

Nell'esempio presentato è stata predisposta una barra di navigazione standard collocata in basso al centro della pagina. L'utilizzo di una barra di navigazione tipo può essere molto utile al fine di predisporre i collegamenti ed i link necessari a creare in modo agevole un *e-book* ed anche per contenere i tempi di realizzazione del prodotto, almeno per quanto attiene le pubblicazioni off-line (CD-Rom/ Floppy). A tale proposito è stata immaginata un barra che contiene delle immagini intuitive facenti parte ormai del bagaglio di conoscenza di un utilizzatore medio di PC, collegate a comandi che consentono all'utente di eseguire funzioni sia elementari che avanzate (figura barra di navigazione). E' possibile inoltre predisporre le icone in modo cioè che appaia una stringa di testo descrittiva associata al puntatore del mouse (passaggio, pressione singola o pressione doppia) e con cambi di colore o di icona.
Nel dettaglio:

a) pulsante icona freccia sinistra: consente di tornare alla pagina precedente;
b) pulsante icona freccia destra: consente di andare alla pagina successiva;
c) pulsante icona lente: consente di avviare la funzione del motore di ricerca, che può essere collegata ad indici full-text creati in precedenza;
d) pulsante icona libro: consente la stampa di parte o di tutto il documento PDF;
e) pulsante icona schedario: consente di selezionare una pagina all'interno del documento alla quale si vuole accedere;
f) pulsante icona posta: consente di aprire una pagina del Web (in questo caso specifico il sito dell'ISTAT) anche se potrebbe essere collegata ad un indirizzo di posta elettronica o ad una qualsiasi pagina del Web;
g) pulsante icona X: consente di uscire dalla pubblicazione elettronica;
h) pulsante giallo ocra (posto in alto a sinistra): consente di richiamare la funzione di Zoom;
i) pulsante icona nota: si tratta di una nota di testo che, una volta attivata con il doppio clic, consente di leggere, in questo caso, informazioni in merito agli autori della pubblicazione elettronica.

L'input in editor per il Web

I programmi di editor per il Web sono simili sia per funzionamento sia per le potenzialità a quelli utilizzati per l'impaginazione professionale. Tuttavia, essendo stati progettati per esigenze diverse da quelle necessarie per le pubblicazioni su carta presentano alcune limitazioni rispetto a queste ultime. Infatti, mentre i file elaborati con programmi di impaginazione professionale possono essere agevolmente trasformati fornendoli di accorgimenti grafici adatti alla pubblicazione su Web è più complesso operare un procedimento inverso.
Inoltre, per le pubblicazioni su Web uno dei requisiti fondamentali è la fruibilità da parte dell'utente del prodotto in tempi brevi e quindi la leggerezza del file elaborato per la sua visualizzazione o il suo scaricamento tramite Internet. Questo rende necessario l'utilizzo di immagini in modalità RGB e in formati "light" ossia a bassa risoluzione. Vengono infatti predilette le immagini che recano il suffisso Jpeg, Jpg, bitmap, eccetera. Inoltre, la "costruzione" di un file tramite l'uso di un software per la pubblicazione su Web composto di link, brevi note di testo e di un buon numero di immagini, ha una forte connotazione ipertestuale che tradotta su carta perderebbe questa sua peculiarità e sarebbe di difficile fruizione. Viceversa, tali software sarebbero utilissimi se utilizzati per le pubblicazioni su supporti magnetici come Floppy Disk e CD-Rom.
Tuttavia se si presentasse la necessità di stampare in Offset i file nati con questo tipo di programmi, occorrerà procedere, durante la trasformazione in PDF, come per i file nati con programmi di video scrittura, avendo l'accortezza, anche in questo caso, di aprire l'immagine/i da inserire nel file costruito con un editor per il Web, portarla/e da modalità RGB in modalità CMKY e salvarla/e in PDF.

L'output confezionato in pacchetti di installazione tramite compilatori

Una volta ottenuto il file PDF partendo da programmi per l'impaginazione professionale (*vedi nota tecnica 2*) o da editor per il Web è possibile "confezionarlo" in maschere o "Form" insieme ad altre applicazioni o file in un pacchetto di installazione tramite dei compilatori, scrivendo qualche riga di codice di programma. Infatti, per la distribuzione su CD-Rom o su Floppy Disk è possibile creare delle applicazioni, ad esempio in Visual Basic, che consentono, non solo di consultare uno o più documenti in formato PDF tramite un'interfaccia amichevole e personalizzata, ma anche di distribuire file aventi un suffisso diverso dal ".PDF".
Un programma così concepito permetterebbe, ad esempio, di inserire tanti *Form* o maschere quante sono le pubblicazioni elettroniche che si vogliono divulgare in un'unica applicazione, creando, in tal modo, per collane o per annate o serie storiche, pubblicazioni elettroniche "complete" consultabili e "navigabili" da un solo CD-Rom. Per quest'ultimo step occorrono delle professionalità esperte nella programmazione e nella progettazione grafica.

Note tecniche

NOTA 1

Installazione di Acrobat Reader

Il programma Acrobat Reader 4.0 richiede, per funzionare correttamente, i seguenti requisiti di sistema:

a) Processore 80486 DX o superiore;
b) Sistema operativo Windows 95/98/NT;
c) Almeno 8 Mbyte di memoria;
d) Circa 3,5/8 MB di spazio su disco fisso;
e) Scheda grafica Svga

Per installare il programma è sufficiente cliccare due volte sull'apposita icona rossa presente nella casella sopracitata. Questa avvia il programma di installazione (" setup") di Adobe. Bisogna quindi leggere le informazioni sulla licenza d'uso, e dopo averle accettate procedere all'installazione.

Il programma è completamente freeware, cioè è possibile distribuirlo senza licenze né attraverso il pagamento di diritti.

I file eseguibili del programma verranno installati nella cartella selezionata, che normalmente è rappresentata da C:\ACROREAD, occupando circa 3,5 MB di spazio su disco fisso.

L'installazione assegna automaticamente il suffisso ".PDF" al programma Acrobat Reader. Ciò significa che facendo due volte "clic" dopo aver selezionato con il puntatore l'icona del file, o dal "Gestione Risorse" di Windows 95, nel caso si usi questo sistema operativo, sarà possibile visualizzare file di testo in formato .PDF, il cui nome termini con questo suffisso.

Questo formato è considerato "universale", in quanto può essere facilmente trasferito su varie piattaforme (anche Macintosh) e visualizzato con gli appositi programmi di rilettura (" reader"). Degli esempi sono collocati nella stessa cartella di installazione del programma: basta localizzarli e
caricarli per apprezzare l'efficienza del programma.

I file possono anche essere stampati, se l'autore del file ha reso possibile questa operazione, in alta risoluzione, cioè con ottima qualità, particolarmente se si usano unità laser o a getto di inchiostro.

Se si dispone di una versione precedente alla 4.0 del programma Acrobat Reader, è consigliabile cancellarla completamente, usando il File Manager od il Gestione Risorse per eliminare la cartella ACROREAD.

La versione di Acrobat 4.0 per Windows 95/98 si disinstalla con la procedura standard, ossia:

"Menu Avvio, Impostazioni, Pannello di controllo, Installazione Applicazioni". Nella lista che appare, si seleziona la voce "Adobe Acrobat Reader", poi si "clicca" il pulsante <Aggiungi/Rimuovi>. Questa operazione cancella tutti i file del programma, le icone nel menu Avvio e le modifiche apportate a Windows durante l'installazione.

Ovviamente, le prestazioni dell'Acrobat Reader sono influenzate notevolmente dalla dimensione dei file PDF che si caricano, dalla quantità di memoria RAM presente e naturalmente dal processore installato nel sistema. E' consigliabile, per un funzionamento ottimale, un sistema di classe Pentium, con almeno 16 MB di RAM, scheda video SVGA accelerata e un modo video dotato di almeno 32.000 colori, meglio ancora se dotato di 16 milioni di colori.

NOTA 2

Modalità per la trasformazione in PDF

Per ottenere la separazione dei colori, operazione necessaria per la stampa tipografica in quadricromia, la trasformazione del file originale in file PDF deve essere effettuata con Acrobat Distiller e non con Acrobat Writer. Infatti, PDFWriter converte i file in modo più rapido ma la risoluzione dell'immagine che produce non fornisce una qualità accettabile per la stampa in Offset. Viceversa, Acrobat Distiller consente di convertire i file PostScript in file PDF rendendo possibile selezionare i font da incorporare nel nuovo file PDF, definire le modalità di compressione delle immagini che vi sono contenute, scegliere un profilo per impostare e separare i colori delle immagini e personalizzare la conversione in numerosi altri modi, consentendo di controllare meglio il processo di conversione e producendo risultati qualitativamente migliori. Esso, infatti, interpreta e converte file PostScript e EPS (Encapsulated Post Script) effettuando una compressione che, rispetto al PostScript, può essere anche 10 o 20 volte inferiore, pertanto va utilizzato per tutti quei documenti che contengono informazioni sulla gestione dei colori o che utilizzano colori indipendenti dal dispositivo. Inoltre, è necessario conservare le caratteristiche PostScript se i file devono essere inviati a un centro di prestampa o ad un laboratorio professionale per la stampa di alta qualità, oppure se si desidera utilizzare un processo batch per convertire i file PostScript in PDF o se si vogliono incorporare font asiatici nel file stesso.
Nelle opzioni di stampa del Distiller è opportuno selezionare la stampa PostScript e selezionare l'opzione "monocromatica" per file originali in Word, in Excel e in ambiente Windows in genere. Inoltre, nella stampa su file di documenti originali redatti in Quark XPress, visto che il programma lo consente, occorre, oltre a operare come sopra specificato, dare la separazione dei colori e stampare su file anche i crocini di registro per consentire al fotolitista un più agevole montaggio delle pellicole.
In generale PDFWriter converte i file in modo più rapido. Mentre Distiller consente di controllare meglio il processo di conversione e produce risultati qualitativamente migliori.
Non tutte le tipografie sono attrezzate per l'uscita in pellicola di file PDF occorre, pertanto, che la società stampatrice sia munita di un apposito RIP (Raster Image Processor), cioè un Computer Server con scheda dedicata o un Software adatto a tale tipo di operazione. Uno di essi è denominato PostScript Level 3 che consente agevolmente di separare i colori da file PDF.

L'input in programmi di video scrittura e fogli elettronici

Nella trasformazione di file in PDF originati con programmi di video scrittura o fogli elettronici e destinati alla stampa Offset bisognerà tenere presente alcune indispensabili operazioni e avvalersi di altri programmi oltre quello di input.
Non tutte le tipografie sono attrezzate per l'uscita in pellicola di file PDF occorre, pertanto, che la società stampatrice che avrà il compito di stampare in Offset il file PDF sia munita di un apposito RIP (Raster Image Processor), cioè un Computer Server con scheda dedicata o un Software adatto a tale tipo di operazione. Uno di essi è denominato PostScript Level 3 che consente agevolmente di separare i colori da file PDF. Comunque, tutte le fotounità di ultima generazione sono in grado di gestire la stampa di file PDF.
Come è stato accennato precedentemente, la separazione dei colori in quadricromia del file (CMYK) Cyan, Magenta, Yellow e Black è indispensabile per stampare in Offset. I programmi di video scrittura non consentono però di operare tale separazione non essendo software dedicati alla stampa di tipo professionale e utilizzando spesso, per la riproduzioni dei colori, la modalità RGB cioè Red, Green e Blu.
Per ovviare a questo problema, nelle opzioni di stampa del Distiller è opportuno selezionare la stampa PostScript Level 3 e selezionare l'opzione "monocromatica". Inoltre è da evitare, pena l'impossibilità di gestire le operazioni di prestampa, di inserire all'interno del file immagini in . jpeg, .jpg, .ico, eccetera ma occorrerà avvalersi di un editor di fotoritocco professionale per aprire l'immagine da inserire nel file di video scrittura, portarla da modalità RGB in modalità CMYK e salvarla in PDF in quanto, come è noto, i programmi di video scrittura non gestiscono immagini ad alta risoluzioni. Infine, occorre copiare e incollare l'immagine con il nuovo suffisso (PDF), posizionandola nel punto desiderato, nel file di video scrittura. Questo è uno dei diversi metodi usati affinché il RIP utilizzato dalla fotolitografia non incontri impedimenti nell'uscita in pellicola della pubblicazione in PDF.

L'input in programmi adatti all'impaginazione professionale

Gli inconvenienti che si riscontrano nella stampa in Offset di file che hanno per input programmi di video scrittura e fogli elettronici non si riscontrano, invece, in file elaborati con programmi di impaginazione professionale. Infatti, tali software elaborano il formato fotocomposto in versione finale già pronto per la stampa in offset e consentono una formattazione di tipo professionale, la separazione dei colori, la gestione di immagini in molti formati ad alta risoluzione e la stampa dei crocini di registro, utili al montatore in fase di assemblaggio delle pellicole. Inoltre, intervenendo sulla/e pagine Mastro del documento è possibile inserire agevolmente, in un secondo tempo, "barre di navigazione tipo" e accorgimenti grafici particolari in modo da rendere la pubblicazione adatta sia alla stampa tradizionale sia alla diffusione su supporti informatici che su Web. Chiaramente, la professionalità richiesta, per la gestione di questo tipo di elaborazione, è elevata.

L'input in editor di immagini e fotoritocco professionale

L'impiego di programmi di editor per il fotoritocco professionale sono indispensabili per tutte le pubblicazioni corredate da immagini. Essi permettono di gestire molteplici formati di immagini e consentono di ritoccarle, modificarle e salvarle in formato PDF, autonomamente da Acrobat. Possono essere impiegati sia per coadiuvare file elaborati con programmi di video scrittura sia con programmi per l'impaginazione professionale che per quelli adoperati per il Web publishing. Anche in questo caso, l'impiego di questi prodotti richiede una formazione specifica.

WORKSHOP

Editoria elettronica per la diffusione dell'informazione statistica

Editoria elettronica vs. cartacea: caratteristiche a confronto

M. Giacobbe

V Conferenza Nazionale di Statistica

Workshop Editoria elettronica per la diffusione dell'informazione statistica

EDITORIA ELETTRONICA VS. CARATTERISTICHE A CONFRONTO

Per la diffusione integrata dell'informazione statistica

di *Marina Giacobbe*

La maturazione delle tecnologie informatiche, e in particolare di quelle telematiche, è una realtà consolidata.

Per una parte degli utenti dell'informazione statistica la facilità di accesso rappresentata da questo canale è sicuramente *irrinunciabile* e l'Istat ha investito - e sta investendo ampiamente - intelligenza e risorse nello sviluppo dell'editoria elettronica.

Accanto a questa, le pubblicazioni cartacee continuano parallelamente la loro vita.

Oggetto di questo intervento sarà il confronto di alcune caratteristiche comunicazionali dei due canali (editoria elettronica /editoria cartacea) rispetto a due elementi:

- i *destinatari* - intesi nel senso più ampio e istituzionale del termine, ovvero nel significato-limite di "tutti i cittadini"
- i *contenuti* - ovvero l'informazione statistica che ci proponiamo di diffondere.

1. I DESTINATARI

Due parole in generale sui destinatari. Chi sono i nostri interlocutori?

1.1 La "carta costituzionale" del sistema statistico nazionale ce li indica molto chiaramente quando dice che i dati elaborati nell'ambito del programma statistico nazionale "sono patrimonio della collettività"; più oltre ribadisce che "enti od organismi pubblici, persone giuridiche, società, associazioni e singoli cittadini hanno il diritto di accedere ai dati" (D.Lgs. 322/89, art. 10, commi 1 e 4).

Secondo questa indicazione, è dunque obiettivo della diffusione che *l'informazione statistica - con il valore di conoscenza e di orientamento che le è proprio - sia reso disponibile a tutti i cittadini*.

Cittadini che sono, nella loro concreta realtà, molto diversi tra loro per caratteristiche personali, condizioni sociali e disponibilità economiche, cultura, livelli di istruzione.
Ma - gli statistici lo sanno meglio di tutti - la diversità non significa affatto disordine: nelle popolazioni alcune caratteristiche si accorpano, si dimostrano connesse, aiutano a costruire fisionomie e profili di gruppo ben chiari e dal forte valore euristico.
Ed è proprio questa - ad avviso di chi scrive - la prospettiva più utile per affrontare la questione della diffusione delle informazioni statistiche e più in generale della *cultura* statistica:

- il tentativo di identificare e connettere in modo organico le caratteristiche salienti del nostro "pubblico", con le diverse esigenze di informazione statistica, allo scopo di rendere disponibili prodotti e servizi che rispondano nel modo migliore alle esigenze della collettività.

O per essere più precisi, dei diversi segmenti di utenza in cui "i cittadini" in generale possono essere raggruppati, costruendo dei *profili di utente* che ci guidino nella progettazione dei prodotti e servizi di diffusione.
Inoltre il tema dei profili di utente è cruciale non solo per la possibilità di rispondere adeguatamente a diversificate esigenze di informazione statistica, ma anche per riuscire a raggiungere e a catturare l'interesse di *quanti non sono utilizzatori* di informazione statistica.

- I prodotti e i servizi offerti dall'Istat sono già oggi diversi e devono esserlo *anche* per avere la capacità di andare incontro alle domande inespresse o finora inconsapevoli della collettività.

1.2 Dal punto di vista del destinatario le caratteristiche dei due canali di diffusione (elettronico vs. cartaceo) non sono simmetriche. In particolare, bisogna soffermarsi su tre temi:

A. La diffusione. Il canale elettronico sconta una importante barriera di accesso, costituita dalla porzione di *popolazione che non ha accesso a computer* e più in particolare ai collegamenti in rete.
E' vero che la diffusione dei computer comincia a raggiungere oggi una penetrazione discreta, ma è tuttavia ben lontana dall'essere pervasiva: secondo i dati dell'indagine Multiscopo del

1999 (beni durevoli posseduti dalle famiglie) il 20,9% delle famiglie italiane possiede un computer. E' vero che agli utilizzatori "da casa" possiamo aggiungere tutti coloro che dispongono di computer sul luogo di lavoro, ma tuttavia non sembra corretto a chi scrive ignorare il tema dell'esclusione di coloro che *non hanno* o *non sanno* utilizzare una opportunità.

Ci sono inoltre notevolissime differenze tra off-line e on-line. E' evidente a tutti la diversa potenzialità rappresentata da un collegamento telematico rispetto all'utilizzazione di un dischetto. In termini di diffusione, ciò significa ridurre ancora, anche solo dal punto di vista puramente quantitativo, la fascia di utenti potenziali.
Secondo gli stessi dati della Multiscopo, nel 1999 solo il 7,6% delle famiglie utilizza il collegamento a Internet. Sia pur considerando, anche in questo caso, gli utilizzatori sul luogo di lavoro e la veloce dinamica di questo fenomeno, il numero di cui si parla si attesta comunque sui 5-6 milioni di persone. Il che significa - in ogni modo - escludere circa il 90% della popolazione già *a monte del tipo di utilizzazione* che effettivamente si fa di Internet.

B. Linearità vs. "spazialità". L'editoria tradizionale (la pagina scritta) risponde ad una logica lineare: qualsiasi "testo" ha uno sviluppo sequenziale, e in questo modo viene affrontato, per consolidata abitudine, dai destinatari.
Viceversa, come tutti sappiamo, la logica dell'informazione digitale è a più dimensioni: la metafora di un "testo" elettronico non può essere che lo *spazio tridimensionale*, in cui il "percorso di lettura" *non è precostituito* e viene definito dall'utilizzatore in base ai propri criteri (ovviamente, entro i limiti degli specifici strumenti a disposizione).
Questo è così evidente che sembra banale ricordarlo in realtà quali Internet, ma è utile avere presente che è questa la logica intrinseca di tutto il canale.

Anche questa differenza ha conseguenze notevoli dal punto di vista dei destinatari; la più significativa è forse quella relativa alla*enciclopedia*, cioè alla quantità e al tipo di nozioni che chi decifra un messaggio deve mettere in gioco per riuscire nel suo intento.
Lo scritto segue la logica lineare - una logica comune, familiare a tutti in quanto appartiene anche al linguaggio parlato - e necessita, come prerequisito, solo del possesso del codice utilizzato (saper leggere, anche se questa capacità può essere definita a vari livelli).

Il canale informatico utilizza meccanismi che sono più simili a quelli che vengono messi in opera da chi deve decifrare una mappa: per farlo bisogna non solo saper leggere, ma riconoscere i segnali grafici, identificare i percorsi e, non ultimo, sapere in che "terra" ci si trova e come ci si arriva quando se ne è ancora lontani.

C. I supporti materiali. Ovvero, carta contro chips, pesanti volumi contro enormi quantità di informazioni disponibili in spazi ridottissimi e a volte del tutto virtuali. E' ovvio che non c'è paragone circa le opportunità di

- *completezza*,
- *facilità di accesso* (per chi sa leggere la mappa),
- *tempestività*

offerte dall'editoria elettronica rispetto a quella tradizionale.

Vediamo la cosa ancora una volta dal punto di vista dei destinatari: avere il computer al posto di una biblioteca, lo schermo invece della carta. E' senz'altro più comodo il primo.
E' *sempre* più comodo il primo?

2. I CONTENUTI

Con questa domanda si entra nel vivo del secondo aspetto della nostra questione, perché la risposta a questa domanda, ad avviso di chi scrive, è proprio: "dipende dai contenuti".

2.1 Ragioniamo sul nostro oggetto istituzionale: *l'informazione statistica ufficiale*. E' fatta in realtà di tanti oggetti diversi, di tanti "pezzi", che possono (o no) diventare suscettibili di diffusione, essere veicolati come prodotti e/o servizi.
Ci sono i dati: assoluti, a diversi livelli e con diversi tipi di aggregazione, su argomenti differentissimi, e poi i metadati, le metodologie, le analisi, le ricerche,.....

- E' chiaro che *non tutto è per tutti*. Ogni tipo di informazione statistica concretizza particolari contenuti, che rappresentano la risposta a specifiche esigenze informative, ed ha pertanto dei destinatari elettivi.

Anche rispetto al tema dei contenuti ritorna in evidenza il discorso dei profili di utente; per esemplificare in modo schematico, diremo che profili di utenza statistica molto specializzati e tecnici avranno bisogno di "contenuti" (e relativi prodotti) altrettanto tecnici e specializzati, e che tipologie progressivamente meno professionali potrebbero avere man mano esigenze più generali, diverse tra loro, motivate da scopi differenti, connesse a "stili" di comunicazione legati alle specifiche contingenze.

2.2 Vorremmo qui sottoporre alla riflessione una questione apparentemente di dettaglio, ma a nostro avviso particolarmente delicata: ad essere precisi, il discorso sulla diffusione andrebbe calibrato in modo ancora più sottile, *non sui profili di utente, ma addirittura sui profili di* ***utenza***.

Vale a dire che non si dovrebbe pensare il nostro destinatario in termini di *persona*-utente, ma bensì in termini di *occasione*-utente: ovvero, la stessa persona fisica potrebbe configurarsi, rispetto a diversi contenuti di informazione statistica, ora come utilizzatore professionale, ora come interlocutore curioso, ora come semplice "orecchiante" che ha necessità di conoscere un dato globale per una semplice utilizzazione empirica.

Se questo è vero, ne seguirà che i tipi di contenuti adeguati alle esigenze dell'occasione-utente potrebbero richiedere strumenti di comunicazione (canali, prodotti) governati da specifiche caratteristiche comunicazionali. Ad esempio:

- quando ho bisogno di un'ampia serie di dati di dettaglio, vorrei disporne collegandomi ad una banca dati;
- quando mi interessa leggere un articolo che analizza un fenomeno, vorrei farlo sulla carta e non sullo schermo del computer;
- quando devo adeguare il canone di affitto, vorrei sapere il tasso di inflazione dal telegiornale.

Così, l'utente - in base alle sue esigenze pratiche, agli usi consolidati, ai pregiudizi, alla possibilità/impossibilità di accesso, eccetera - selezionerà il canale di comunicazione cui rivolgersi: questo significa che il nesso "tipo di contenuto-canale di trasmissione" può essere considerato particolarmente forte e pertinente.

Quanto sopra è detto per sottolineare ancora come sia necessario

- prevedere la massima flessibilità nel progettare i prodotti e servizi statistici, ed insieme prevedere anche la loro massima *integrazione*.

L'obiettivo da proporsi è infatti quello che tutte le opportunità di informazione statistica costituiscano un *sistema*, e risultino in questo modo il più esaurienti possibile nei confronti delle esigenze della collettività, nelle sue diverse articolazioni.
Il canale telematico risulterà ad esempio irrinunciabile per le utilizzazioni avanzate e per alcune professionalità di riferimento (es. manager, giornalisti), mentre è senz'altro auspicabile *affiancare* alle opportunità tecnologiche i canali dello scritto per contenuti meno specialistici - e in particolare per la diffusione di dati corredati da analisi.

2.3 Non solo: riflettiamo ancora una volta su coloro che *non sono* utilizzatori di informazione statistica: destinatari anch'essi, secondo i nostri fini istituzionali, ma verso i quali il nostro atto comunicativo è fino ad ora andato disperso.
Proprio verso questo pubblico è essenziale trovare delle forme di comunicazione attraenti, amichevoli, che "prendano per mano" le persone attraverso i loro interessi e abitudini. E farlo attraverso il canale più adeguato è un elemento cruciale nell'equilibrio della comunicazione.

3. IMMAGINE PER UNA SINTESI

La questione si configura quindi dal punto di vista della diffusione in questo modo: quali contenuti trasmettere, verso quali destinatari, servendosi di quali canali di trasmissione?

3.1 Una proposta di sintesi: possiamo immaginare

- i *destinatari* (da quelli più generici a quelli più tecnici e professionali),
- i *contenuti* (da informazioni statistiche "banali" a files manipolabili per specifiche esigenze di ricerca)
- i *canali* (eletronico vs scritto tradizionale vs parlato...)

come tre *dimensioni* che disegnano un "oggetto" (il nostro prodotto o servizio di informazione statistica). Solo se questo ha tutte le "dimensioni" - cioè le caratteristiche - giuste per collocarsi nello spazio che gli è destinato - l'esigenza dell'utente - il nostro atto comunicativo avrà successo.
E' infatti convinzione di chi scrive che trovare le aree di coerenza (*tridimensionali*), le connessioni *forti* tra tipi di utenti, tipi di informazioni e canali utilizzati costituisca

1. *un aspetto irrinunciabile* per il buon esito della comunicazione: perché solo la possibilità di trovare - per così dire - "ciò che mi serve" (cioè l'informazione statistica, il contenuto) "nel modo che mi serve" (attraverso il prodotto o il servizio che so/posso utilizzare), secondo il mio specifico fine e secondo le mie caratteristiche, rappresenterà il felice esito della comunicazione;
2. *una via privilegiata* per l'allargamento della diffusione dell'informazione statistica: perché la possibilità di disporre di informazione statistica in modo amichevole e accattivante, strettamente legato alle occasioni, agli interessi, alle abitudini delle persone, può rappresentare una via di accesso potente per trovare ascolto, accostare nuovo pubblico all'utilizzazione delle informazioni e con ciò contribuire - col tempo - a diffondere una migliore cultura statistica nel paese.

3.2 In conclusione, si propongono qui come possibile oggetto di riflessione le seguenti considerazioni:

1. Che il canale elettronico, ed in particolare telematico - benché si configuri a sua volta secondo diverse articolazioni ed abbia guadagnato nel suo complesso moltissimo terreno in termini di fruizione e di amichevolezza - si configura a tutt'oggi come canale di uso parziale, fruibile dal segmento di utenza che può definirsi a vario titolo "avanzata" (ricercatori, manager e professionisti, *giovani*); d'altronde, proprio per questo specifico pubblico la telematica offre, rispetto ai dati, opportunità di completezza, possibilità di approfondimento e insieme semplicità di accesso, tempestività che sono semplicemente impensabili per qualsiasi altro mezzo di trasmissione dell'informazione.

Il telematico può essere il canale privilegiato del "dato" (dei MILIONI di dati: è il luogo dell'esaustività, in cui ogni singolo segmento di informazione statistica viene registrato e reso reperibile).

2. Che lo scritto (cioè la diffusione tramite l'editoria tradizionale) si potrebbe considerare come strumento particolarmente adeguato (e dedicabile?) ad un pubblico che si potrebbe definire "intermedio": utenti che non sono utilizzatori professionali della statistica, ma che *possono esservi* in qualche modo connessi (imprenditori, funzionari pubblici, professionisti); persone che per interesse di cultura generale o di specifiche problematiche si accostino ad un tema in prospettiva statistica (studenti, insegnanti, curiosi e lettori di ogni genere); più largo pubblico da informare su specifici argomenti e/o dati.
Tutti costoro si potrebbero considerare *accomunati da un tratto saliente*: l'esigenza di poter trovare - oltre e insieme dato - un'indicazione di lettura dello stesso, la possibilità di accompagnare il numero o la serie di numeri con una griglia di analisi che consenta al non-specialista di *trasformare il dato in informazione*. In questo senso, lo scritto - nelle sue diverse articolazioni e registri, dal saggio alla divulgazione - potrebbe configurarsi come il canale privilegiato per l'obiettivo di diffondere l'informazione statistica quale strumento di analisi dei fenomeni.
E' inoltre possibile un ruolo - forse importante, comunque su cui riflettere - per iniziative di comunicazione non-mediata (pubblicazioni divulgative e altro).

Lo scritto può essere il canale della "lettura del mondo" (dei dati legati al loro SIGNIFICATO: è il luogo in cui essi servono un percorso di comprensione della realtà - di una specifica realtà).

3. Infine che il parlato - il canale di comunicazione primario, rivolto a tutti e privo in linea teorica di qualsiasi barriera di accesso - è uno strumento potentissimo e insieme labilissimo. Alla possibilità di raggiungere - attraverso radio e TV - grandi quantità di persone si affianca la questione della durata *istantanea* della comunicazione.
In questi casi, contrariamente ai precedenti, il dato non vive nella relazione diretta con il fruitore, ma passa attraverso una persona concreta, e cioè colui che lo presenta e il suo modo

di proporsi. Perciò è assolutamente cruciale governare in maniera precisa ogni elemento della comunicazione, per fare in modo che - per così dire - "la statistica buchi lo schermo".
(A margine si rammenta, e solo per citazione, un argomento, che avrebbe bisogno di ben altri approfondimenti. Il telefono rappresenta un mezzo di diffusione individualizzata di diffusione della cultura statistica: il centro diffusione, il call-center, le indagini telefoniche, hanno tutti elementi in comune con quanto detto sopra, e anche specificità comunicazionali sulle quali varrebbe la pena di ragionare).

Il parlato è il canale della "presenza" (il numero in versione FAST FOOD: consumo rapido - e forse non sempre salutare - ma tuttavia, se ben governato, saporoso e capace di "agganciare" un largo pubblico).

3.3 Se tutto ciò è condiviso, potrebbero seguirne alcune indicazioni:

1. Ciascun tipo di contenuti è adatto ad essere diffuso preferibilmente tramite uncanale piuttosto che un altro, in relazione al destinatario prescelto. Identificare correttamente il canale è fattore di successo della comunicazione (cioè di soddisfazione dell'utenza e adempimento di un compito dell'Istituto)

2. Per converso, alcuni tipi di contenuti non sono adatti a determinati canali

3. Nel trasferire contenuti da un canale all'altro è necessario rimodellare i messaggi (perché presumibilmente cambieranno i destinatari e cambia comunque il linguaggio specifico, caratteristico di quel canale).

In conclusione:

La diffusione va pensata come *sistema*, tenendo presente la molteplicità dei fattori implicati (destinatari, contenuti, canali di comunicazione) in modo che ogni segmento abbia una specifica funzionalità e - oltre a soddisfare un bisogno - contribuisca al consolidamento/ miglioramento di tutto l'insieme.

WORKSHOP

Editoria elettronica per la diffusione dell'informazione statistica

La popolazione residente nei comuni italiani

V. Patruno, M. Venturi

Quinta Conferenza nazionale di Statistica:
Sessione: Editoria elettronica per la diffusione dell'informazione statistica

Demo.istat.it: La Popolazione Residente nei Comuni Italiani

Vincenzo Patruno - Istat – DINF RSI/B
Marina Venturi - Istat – DCPT DEM/1

Introduzione

L'evoluzione tecnologica e la diffusione sempre più capillare di Internet in tutti gli ambienti (domestici, aziendali, scientifici e professionali) ha determinato un radicale cambiamento nelle modalità e nelle politiche di diffusione di dati ed informazioni e nella erogazione di servizi in tutti i settori.
In particolare, in campo statistico, dove cresce la necessità da parte degli utenti di avere dati sempre più aggiornati e sempre più dettagliati, la predisposizione di sistemi *online* via Internet consente di raggiungere entrambi gli obiettivi.
Infatti, i guadagni in termini di tempestività sono legati prevalentemente al totale abbattimento dei lunghi tempi necessari per la predisposizione di pubblicazioni su carta. Inoltre, la quantità di dati che è possibile estrarre attraverso il Web è di gran lunga superiore a quella che normalmente può essere oggetto di una pubblicazione.
Proprio per rispondere in tempo reale alla crescente domanda di informazione statistica in materia demografica, e per fornire dati al massimo livello possibile di dettaglio è stato progettato e realizzato il sistema ***Demo – Popolazione Residente nei Comuni italiani*** attualmente operativo all'indirizzohttp://demo.istat.it.
I dati diffusi sono il risultato di tre distinte indagini condotte regolarmente dall'Istat presso gli uffici di Anagrafe dei singoli Comuni:

- **Movimento e calcolo della popolazione residente (Indagine P2)**
  Le informazioni riportate sono:
  - Popolazione residente a inizio anno (ottenuta a partire dal dato dell'ultimo censimento, aggiornato con i saldi naturali e migratori relativi al periodo tra il 20 ottobre 1991 e l'inizio dell'anno considerato)
  - Nati vivi (da genitori residenti nel Comune)
  - Morti (già residenti nel Comune) · Saldo naturale (differenza tra le due voci precedenti)
  - Iscritti per trasferimento di residenza (da altri Comuni o dall'estero)
  - Cancellati per trasferimento di residenza (ad altri Comuni o all'estero)
  - Saldo migratorio (differenza tra le due voci precedenti)
  - Popolazione residente a fine anno (popolazione d'inizio anno + saldo naturale + saldo migratorio)

- **Cittadini stranieri iscritti in Anagrafe: Bilancio Demografico e Popolazione residente per sesso e cittadinanza al 31 dicembre (Indagine P3)**
  Si forniscono le notizie sui flussi annuali naturali (nati vivi, morti) e migratori (iscritti, cancellati) nonché sull'ammontare della popolazione straniera residente alla fine dell'anno, distinta per sesso e cittadinanza con l'indicazione per sesso della quota dei minorenni.

- **Popolazione residente per età, sesso e stato civile (Indagine POSAS)**
  Si riporta la distribuzione della popolazione per sesso, età e stato civile al 1° gennaio dell'ultimo anno disponibile. Poiché la rilevazione non è obbligatoria, i dati non sono disponibili per tutti i Comuni italiani; tuttavia, il livello di collaborazione dei Comuni è andato crescendo nel tempo fino a raggiungere la quota del 98.6%. Le informazioni a livello provinciale ed aggregazioni superiori sono sempre presenti in quanto, in caso di mancata copertura (dato mancante per qualche Comune) si procede ad una stima.

DEMO è un sistema per l'interrogazione dei dati raccolti attraverso le suddette indagini. Consente infatti di scegliere, per ogni rilevazione, uno specifico livello territoriale (L'intera Italia o una Ripartizione una Regione, una Provincia o, addirittura, un singolo Comune) attraverso una selezione di tipo gerarchico e le variabili di interesse.
In particolare, per l'indagine POSAS, sarà possibile scegliere quali tra le modalità che descrivono lo stato civile devono essere visualizzate (ad esempio i soli celibi e nubili) e l'intervallo di età a cui fare riferimento.
Per l'indagine P2 la scelta potrà essere effettuata soltanto sul livello territoriale ed il sistema fornirà la tabella contenente l'ammontare della popolazione e i dati di flusso.
Infine, per l'indagine P3, oltre al livello territoriale sarà possibile selezionare gruppi di cittadinanze. In questo caso, verranno visualizzate due tavole distinte, una contenente la distribuzione della popolazione straniera per sesso relativa alle cittadinanze selezionate ed una che riporta il dato di stock sulla popolazione straniera ad inizio e fine anno e il movimento naturale e migratorio.
Il risultato dell'interrogazione è visualizzato sul *browser* dell'utente ma può essere anche salvato direttamente in formato testo ed essere così utilizzato dall'utente per elaborazioni personali con fogli elettronici sul proprio *Personal Computer*.

L'interfaccia grafica

L'interfaccia è pensata per essere di facile e immediato uso da parte dell'utente.
Sulla home page sono messi bene in evidenza i link che portano alle tre sezioni relative alle indagini sopra menzionate. (Fig. 1). Facendo click col mouse sulla tipologia di indagine di cui si desidera estrarre i dati, si accede ad una nuova pagina che ha la caratteristica di contenere tutti gli strumenti necessari all'utente per selezionare i parametri e le variabili di interesse. La novità dell'applicazione risiede

proprio nell'aver concentrato in una unica pagina tutti i criteri di selezione, contrariamente a quanto avviene oggi nella gran parte delle applicazioniweb che si sviluppano in profondità e in cui, prima di accedere all'informazione desiderata, è necessario effettuare numerosi click col mouse; poiché il numero di click effettuati rappresenta la distanza che separa l'utente dall'informazione desiderata, più click vengono effettuati e più lontana è l'informazione cercata; in conseguenza di ciò, aumenta anche il tempo necessario alla ricerca, oltre che disorientare ed essere dispersivo per l'utente.

Fig. 1

L'aver quindi concentrato la possibilità di effettuare le scelte desiderate in un'unica schermata, ha reso da una parte lo strumento estremamente comodo da utilizzare, dall'altra ha ridotto al minimo i tempi di accesso ai dati. (Fig. 2).
L'interfaccia è stata pertanto costruita utilizzando tre *frames,* due per l'impostazione delle *query* sulla base dati e l'altra per mostrare, sotto forma di tavola, i risultati ottenuti (Fig. 3). L'interazione tra *frames*, la gestione dei parametri scelti e la gestione della gerarchia territoriale sono state realizzate attraverso codice *javascript.*
Di sicuro interesse è la gestione della gerarchia territoriale fino a livello comunale.

Mentre fino a livello di Provincia le dipendenze tra Ripartizioni, Regioni e Province vengono gestite direttamente dalle applicazioni *javascrip*t, i Comuni corrispondenti ad una provincia selezionata vengono "caricati al volo" sul *browser* dell'utente. In sostanza, ciò significa che, una volta selezionata una provincia, parte automaticamente una *query* sul DB che restituisce e visualizza nella *select box* corrispondente tutti i Comuni appartenenti a quella Provincia.
In questo modo, l'utente avrà sul proprio *client* solo i Comuni della Provincia a cui è interessato in quel momento. D'altra parte sarebbe stato impossibile oltre che inefficiente caricare tutti gli 8100 comuni all'inizio della sessione.
Tutte le applicazioni *javascript* sono state realizzate utilizzando le funzionalità supportate da *Internet Explorer 5.x*. Il sistema è stato comunque adattato per poter funzionare anche con il *browser Netscape 4.xx*, in considerazione della sua larga diffusione tra gli utenti.

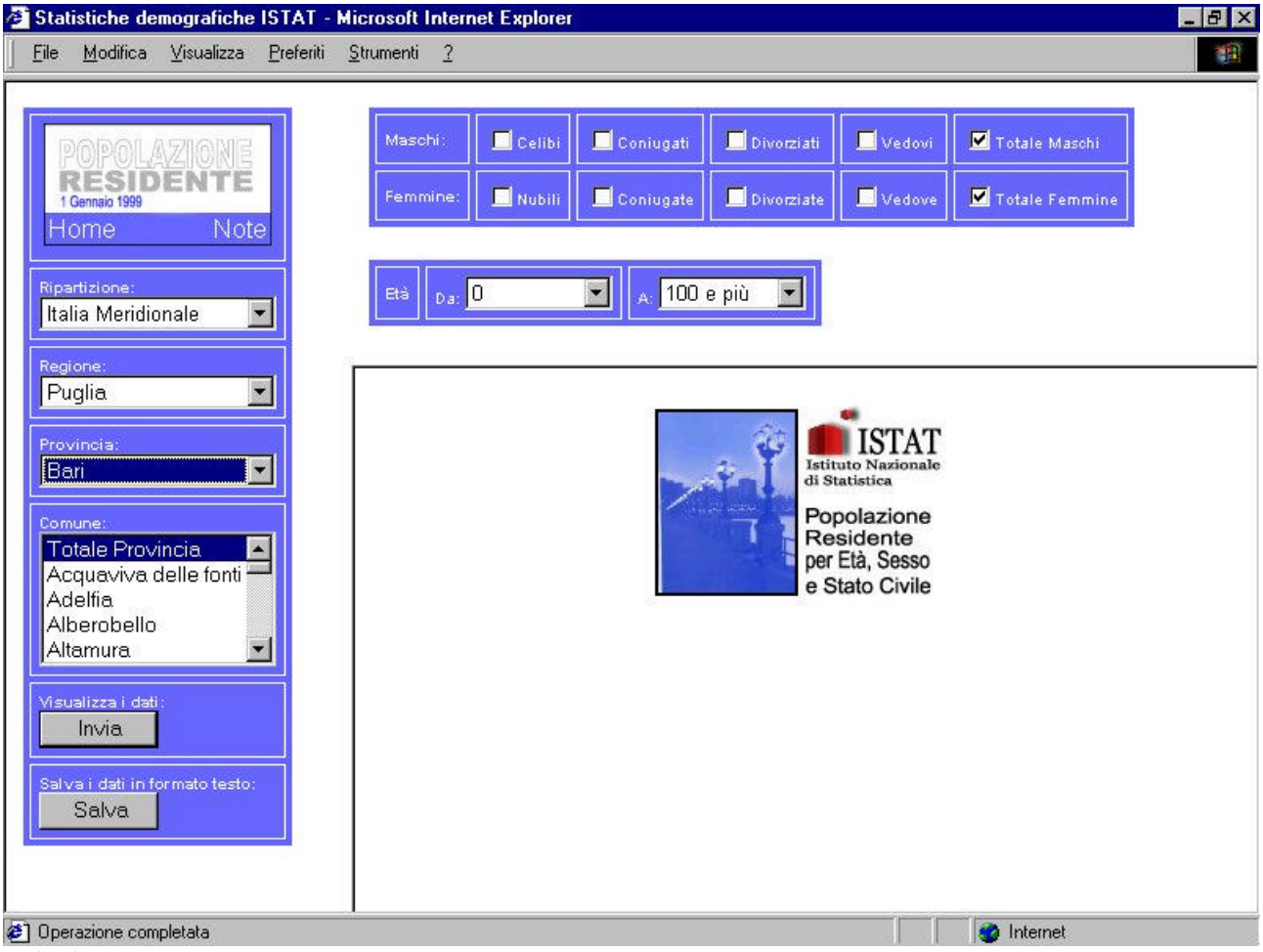
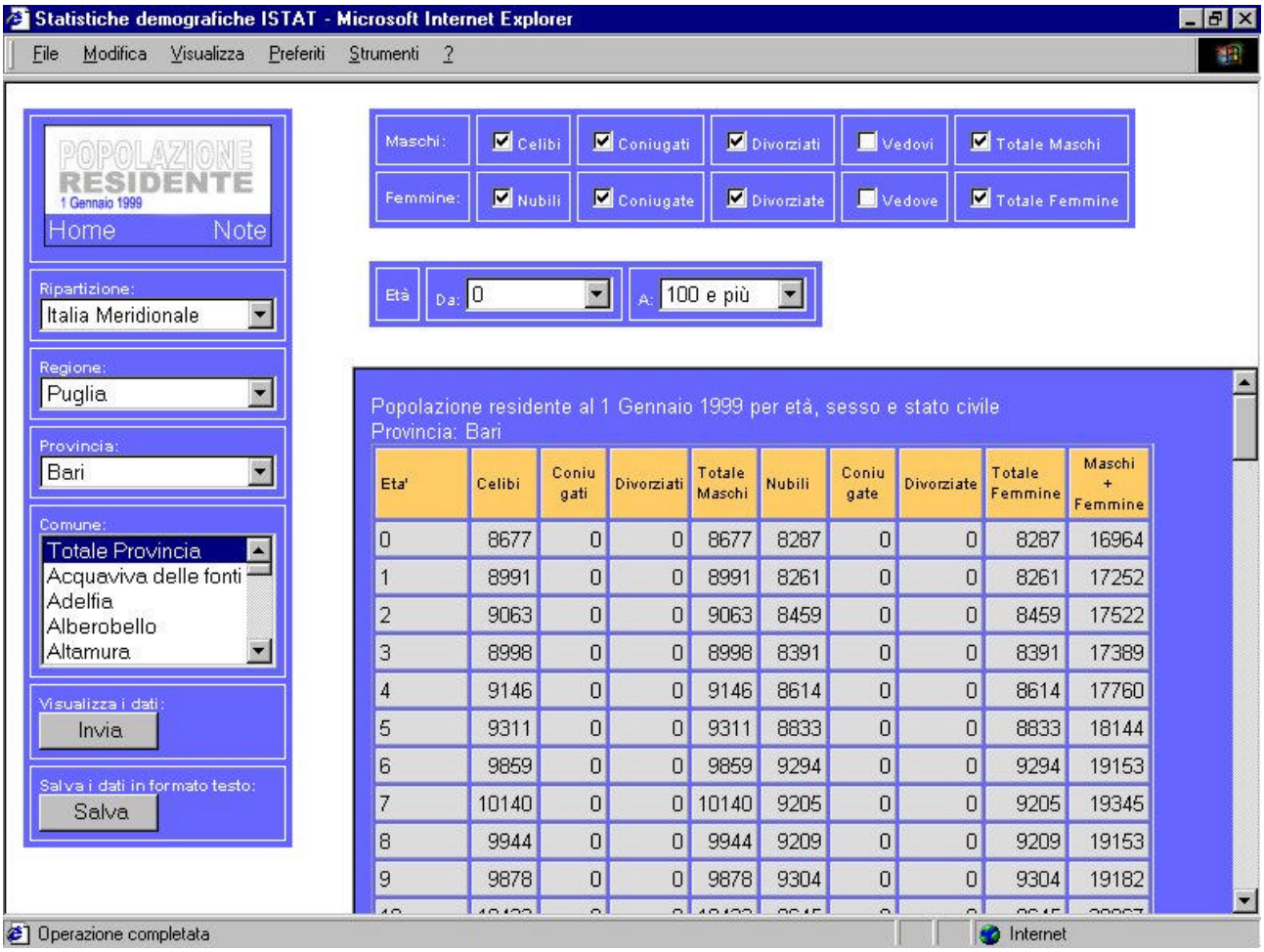


Fig. 2

L'organizzazione dei dati.

I dati sono stati organizzati su un data base relazionale. In realtà, le tabelle sono state costruite in modo tale da contenere non solo i dati elementari al massimo livello di dettaglio, ma anche le diverse aggregazioni corrispondenti ai vari domini territoriali.

In questo senso, il modello di dati assomiglia molto ad un modello multidimensionale tipico dei *data warehouse*. Il *database* risulta quindi ridondante, a beneficio tuttavia delle prestazioni dell'intero sistema, che risulta avere una elevata velocità di risposta alle *query* dell'utente, sebbene i dati complessivi occupano uno spazio di oltre 200 MB.



Eta'	Celibi	Coniugati	Divorziati	Totale Maschi	Nubili	Coniugate	Divorziate	Totale Femmine	Maschi + Femmine
0	8677	0	0	8677	8287	0	0	8287	16964
1	8991	0	0	8991	8261	0	0	8261	17252
2	9063	0	0	9063	8459	0	0	8459	17522
3	8998	0	0	8998	8391	0	0	8391	17389
4	9146	0	0	9146	8614	0	0	8614	17760
5	9311	0	0	9311	8833	0	0	8833	18144
6	9859	0	0	9859	9294	0	0	9294	19153
7	10140	0	0	10140	9205	0	0	9205	19345
8	9944	0	0	9944	9209	0	0	9209	19153
9	9878	0	0	9878	9304	0	0	9304	19182

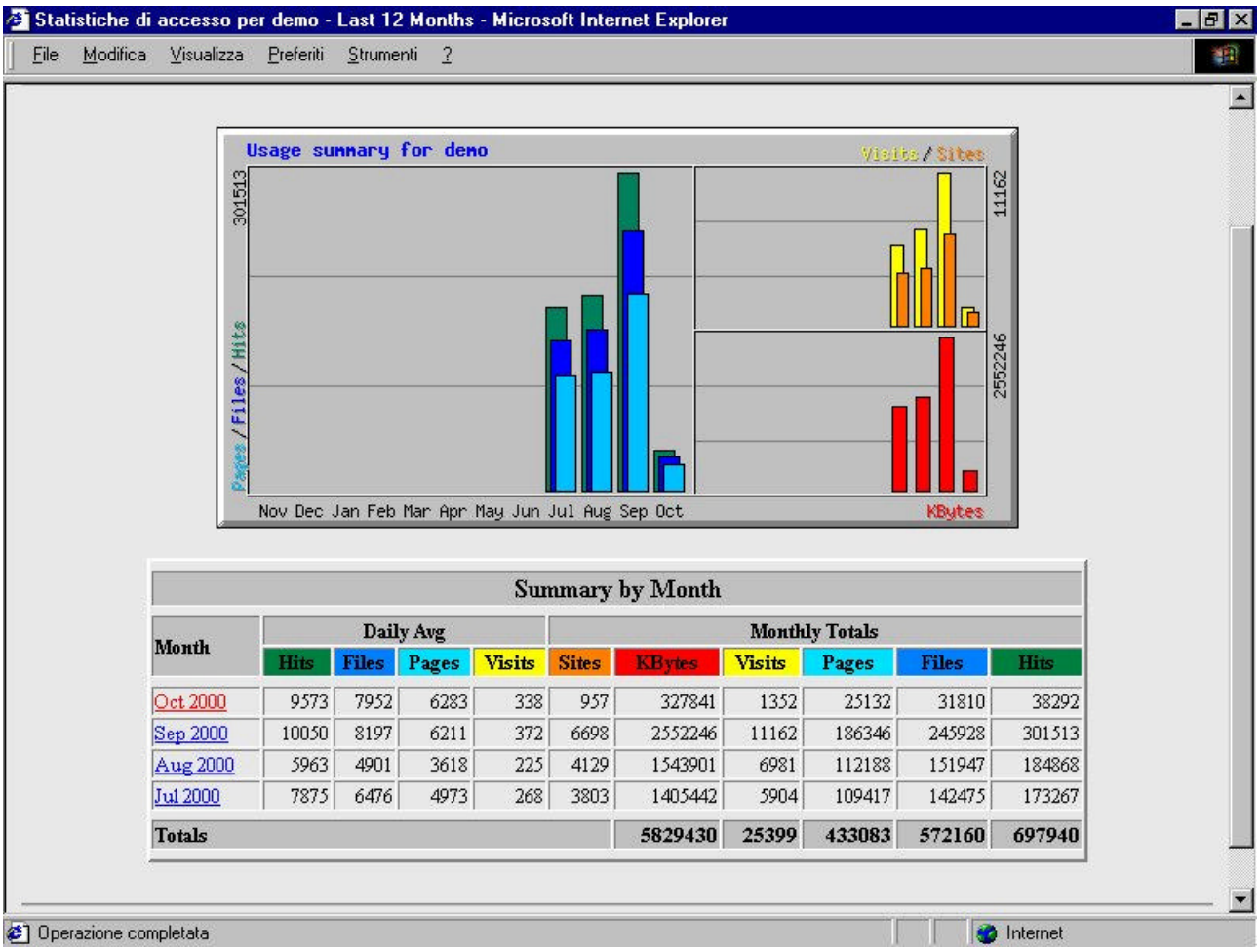
Fig. 3

L'ambiente di sviluppo e la piattaforma HW

Tutto il sistema è stato realizzato utilizzando *script* PHP3. La versatilità del linguaggio e la facilità con cui consente di colloquiare con la base dati, nonché le numerose funzionalità supportate hanno consentito la realizzazione di tutte le applicazioni necessarie in tempi brevi. Il software è stato implementato su una *workstation UNIX-AIX* a garanzia di una alta affidabilità, prestazioni di buon livello nonché di un alto livello di sicurezza.

Monitoraggio degli accessi

Gli accessi all'applicazione sono costantemente monitorati. Ogni giorno vengono elaborate automaticamente le statistiche di accesso. Al momento, considerando che il sistema è attivo dalla fine di Giugno 2000, si hanno a disposizione i dati relativi ai mesi di Luglio, Agosto e Settembre che vengono riportati nella figura seguente (Fig.4).
Tra le varie statistiche riportate, il numero di "visits" rappresenta il numero di sessioni attivate da uno o più utenti. Il numero medio di "visits" giornaliero è stato, nel mese di settembre di 372, pari a circa 11.162 visite mensili, a testimonianza di una elevato interesse all'applicazione, da parte dell'utenza esterna.



Summary by Month

Month	Daily Avg				Monthly Totals					
	Hits	Files	Pages	Visits	Sites	KBytes	Visits	Pages	Files	Hits
Oct 2000	9573	7952	6283	338	957	327841	1352	25132	31810	38292
Sep 2000	10050	8197	6211	372	6698	2552246	11162	186346	245928	301513
Aug 2000	5963	4901	3618	225	4129	1543901	6981	112188	151947	184868
Jul 2000	7875	6476	4973	268	3803	1405442	5904	109417	142475	173267
Totals						**5829430**	**25399**	**433083**	**572160**	**697940**

Fig. 4

Sviluppi futuri

Sulla base delle indicazioni fornite sia dagli utenti interni che da quelli esterni, oltre che dai responsabili delle diverse indagini implementate in Demo, sono previsti a breve degli ampliamenti del sistema che riguardano da una parte l'inserimento di

funzionalità aggiuntive e dall'altra, l'arricchimento della base dati con informazioni di carattere congiunturale.
In particolare è in corso di sviluppo una *utility* per consentire ad un utente lo scarico in formato testo sul proprio Personal Computer dei dati relativi a gruppi di Comuni (tutti quelli appartenenti ad una determinata Provincia), gruppi di Province (tutte quelle appartenenti ad una singola Regione), gruppi di Regioni (tutte quelle appartenenti ad una Ripartizione geografica) ed infine gruppi di Ripartizioni.
Rispetto all'ampliamento della base dati, verranno messi a disposizione anche i dati riguardanti la popolazione residente nelle regioni italiane distinta per sesso ed età al 1° gennaio del corrente anno, ottenuti applicando dei modelli di stima.

RELAZIONE DEL DIRETTORE DEL DIPARTIMENTO DELLA SEGRETERIA CENTRALE DEL SISTEMA STATISTICO NAZIONALE

Vincenzo Lo Moro

direttore di dipartimento dell'Istituto Nazionale di Statistica

1. Introduzione

Fornire un bene pubblico essenziale; fornire il bene "informazione statistica ufficiale"; fornirlo come sistema; fornirlo in una società tecnologicamente avanzata, nella società dell'informazione: vorrei partire da questo livello "alto" per giungere a definire alcuni criteri chiave della nostra azione concreta.

Nella Quarta conferenza di statistica mi sono soffermato su una declinazione della natura pubblica dell'informazione statistica fornita dal Sistema statistico nazionale. La richiamo brevemente:

- assenza di convenienza alla produzione privata;
- non escludibilità e non rivalità nel consumo;
- indefinibilità del valore di scambio;
- non recuperabilità dell'investimento;
- meritorietà, perché eleva il livello di conoscenze e la consapevolezza civile, ovvero riduce l'incertezza informativa;
- equità, perché pone, potenzialmente, tutti i cittadini su un unico piano informativo;
- ampia possibilità di produzione congiunta (ad esempio connessa allo sfruttamento delle informazioni amministrative);

e, infine:
- bene “comune” nel senso che è parte integrante della funzione di produzione di molti altri beni e servizi, anche se non di tutti, in quanto supporto decisionale per organizzare i processi produttivi stessi.

Queste caratteristiche possono essere ricondotte a diversi filoni di riflessione teorica, ma colgono tutte un qualche aspetto presente nell’informazione statistica pubblica. Le prime quattro danno una caratterizzazione “in negativo” e stabiliscono la necessità di una produzione pubblica (chi debba produrre). Le altre quattro, in positivo, indicano quale è la ragione della produzione, la finalità effettiva, che non bisogna mai perdere di vista.

Ogni singola rilevazione statistica presa da sola, e sottoposta al vaglio dei criteri indicati, non necessariamente li supera, specie nella società tecnologica caratterizzata da forte circolazione di informazione. Esistono ormai schemi contrattuali che consentono di escludere dal consumo; ci sono statistiche per loro natura “pubbliche”, con un forte segnale previsionale, come quelle meteorologiche o quelle finanziarie, che possono essere prodotte da privati, perché recuperano i costi di produzione (e anche di investimento) attraverso una politica commerciale aggressiva, grazie anche ad una domanda crescente; ci sono statistiche che non sono più meritorie o che non sono neanche un bene intermedio utile.

Non è un caso che una delle recenti evoluzioni legislative quali il DPCM del 9 marzo 2000 preveda, pur con opportune cautele, l’ingresso di soggetti privati nel Sistan.

Ma a fronte delle critiche che possono essere rivolte alla singola rilevazione (e che vanno effettuate come screening periodico, per non portarsi dietro attività del tutto inutili o meglio effettuabili da altri operatori), nessuno può mettere in discussione che il Sistema statistico nazionale e i subsistemi che lo compongono - sia tematici che territoriali – rispondono fortemente a quelle caratteristiche che abbiamo delineato.

Nessuno può mettere in discussione vuol dire che non ci può essere contestazione. Contestazione non ci può essere che ad una condizione: l’opera di manutenzione e razionalizzazione delle attività del Sistema sia costantemente portata avanti, direi certosinamente, ma anche con ambizione e creatività. Conservando e innovando quel tanto che è necessario per mantenere sempre vive le ragioni, le utilità del sistema, senza paura, ove occorre, di tagliare.

Questo è il lavoro che è stato svolto negli anni trascorsi ed è quello che mi impegno a portare avanti nei prossimi con qualche accelerazione e qualche accento, che vorrei sottoporre all’attenzione di questa Conferenza.

2. Uffici di statistica

Il livello minimo strutturale del Sistema, la cellula primaria è "l'ufficio di statistica".

Sappiamo quanto questi siano diversi tra di loro a seconda del contesto in cui lavorano, ma un corpo non sarebbe tale se tutte le cellule fossero uguali.

Se oggi dovessimo indicare la collocazione organizzativa degli uffici di statistica forse non utilizzeremmo più la terminologia del d.lgs. 322/89 e delle successive direttive del Comstat – "…funzioni organicamente distinte; settore a sé stante; dipendenza dall'organo di governo, etc." – bensì in modo più appropriato semplicemente la dizione – "centro di responsabilità" intesa a garantire a quell'ufficio la tipica autonomia gestionale assicurata a queste strutture previste dalle riforme amministrative degli anni '90. L'Ufficio di statistica definirebbe obiettivi e utilizzerebbe risorse come qualsiasi altra struttura dell'amministrazione, ma con una forte ottica trasversale e di servizio alle altre componenti. Una "carta dei servizi interna" stabilirebbe, inoltre, le competenze e i reciproci obblighi tra le strutture.

L'ufficio di statistica di una amministrazione è oggi impegnato per lo più su due macrofunzioni:

- la collaborazione con altre aree dell'amministrazione;
- la produzione di statistiche, sia per conto dell'amministrazione che nell'ambito del PSN.

3. Il supporto alle altre aree dell'amministrazione

3.1 Uffici di statistica e funzione statistica distribuita

I dirigenti di strutture, i responsabili di procedimento, i responsabili di politiche, di azioni e di progetti, si avvicinano tutti alla statistica, almeno a quella di base. Più precisamente, sono soggetti che debbono trattare misure di risorse e misure di risultato, report periodici, quadri di controllo, indicatori per la loro attività corrente, al fine di compiere valutazioni e fornire elementi per farsi valutare.

Da un lato tutto ciò è previsto dalle norme, dall'altro ognuno ne avverte l'oggettiva necessità per il lavoro corrente. La prassi di crearsi piccoli e grandi sistemi di raccolta di informazioni è sicuramente diffusa, se non altro per avere informazione statistica in tempo reale in caso di improvvisa richiesta da parte dell'organo politico. Nella maggior parte dei casi, peraltro, queste informazioni non escono dall'amministrazione ma tornano utili alla stessa secondo percorsi impossibili da ricostruire. Sovente emergono ridondanze; difficil-

mente si ha a che fare con schemi ben strutturati come vorrebbe l'ortodossia delle tecniche informative.

Questo livello di statistiche distribuite vede impegnato l'ufficio di statistica in un obiettivo comune con l'organizzazione alla quale appartiene, nell'intento di sviluppare la valorizzazione delle informazioni detenute dall'amministrazione stessa a fini gestionali.

La situazione conseguente alla riforma della P.A., in direzione della divisionalizzazione, presenta vantaggi e svantaggi. Da un lato, i centri di responsabilità legati a particolari funzioni tendono a sviluppare propri sistemi informativi di raccolta, proprie unità statistiche e mal sopportano ingerenze di altri settori. Ma, dall'altro, se è vero che debbono fare i conti con un rapporto stringente tra obiettivi assegnati e risorse, l'utilità diventa manifesta a entrambi, il "potere" dell'ufficio di statistica si allarga e l'amministrazione ne beneficia.

E' su questo circolo virtuoso che dobbiamo puntare, anche se lo scenario della qualità degli archivi informativi esistenti, della capacità di trattare le informazioni e della volontà di cooperare è a volte sconfortante.

3.2 Uffici di statistica e controllo interno

La connessione tra queste due funzioni è più volte richiamata dalla normativa sia con riferimento alle amministrazioni centrali che alle altre amministrazioni pubbliche. Le tecniche di controllo interno hanno la loro base in parecchie discipline e si alimentano principalmente di cifre.

Le recenti norme (L.59/97; d.lgs. 286/99) ribadiscono la necessità di costituire un sistema informativo-statistico di supporto all'attività di controllo interno collegando strettamente questa attività a quella dell'ufficio di statistica. Che funzione ha questo collegamento? Non certo, o comunque marginalmente, quella di fornire dati, piuttosto quella di collaborare alla costruzione del sistema informativo-statistico per il controllo, fornire metodi e tecniche, certificare il sistema, condividere alcune informazioni macro che possono avere una valenza anche al di fuori dell'amministrazione (ad esempio per le istanze di controllo esterno). Il Dipartimento della funzione pubblica non ha chiesto all'Istat, come prevedeva il d.lgs. 29/93, di "elaborare norme tecniche e criteri per le rilevazioni ed analisi di costi e rendimenti". Se lo avesse fatto, e finanziato effettivamente, il canale Sistan e il canale controlli interni avrebbe avuto ben più forti connessioni.

Oggi, il legame è più tenue. Ciò non toglie, tuttavia, che esistono momenti di forte interscambio tra le due istanze, molto legati alle esperienze nelle singole amministrazioni e a cui il supporto formativo va garantito con continuità.

Esiste, anche in questo caso, un circolo virtuoso che può essere attivato nella cooperazione interna all'amministrazione, puntando ad un riconoscimento sostanziale del ruolo e della professionalità dell'ufficio di statistica.

3.3 Uffici di statistica e sistemi informativi tematici

Si è già accennato al fatto che i singoli centri di responsabilità tendono a costruire sistemi informativi propri e a chiudersi in un'ottica proprietaria dei dati. Quando le dimensioni e il tema trattato sono particolarmente rilevanti nascono "osservatori, sistemi articolati e integrati di raccolta e trattamento delle informazioni". Dal punto di vista dell'ufficio di statistica, questo può essere vissuto come una espropriazione di competenze, come una concorrenza in cui normalmente si è perdenti. Una prassi consolidata, ma assolutamente da contrastare, è quella per cui il colloquio tra uffici centrali e territoriali privilegia i soggetti tematici, saltando gli uffici di statistica. Se questa si poteva giustificare agli albori del Sistema, oggi diviene destabilizzante.

L'esperienza ci ha insegnato che su questo terreno non sono ammesse scorciatoie: saltare gli uffici di statistica per appoggiarsi esclusivamente sui depositari degli archivi informatizzati può funzionare nel breve periodo, ma alla lunga rivela tutti i suoi limiti. Ricordiamoci che la statistica è anche ripetitiva (in quanto serve anche la documentazione storica) e non vive delle realizzazioni di un solo anno. A presidiare questa esigenza di continuità non possono che esserci gli uffici di statistica.

La costruzione di sistemi informativi tematici, di amministrazione, di territorio o nazionali non può comunque essere considerata una sconfitta. Anzi il Sistan sarà sempre di più un Sistema di sistemi, dei quali sorveglia e coordina poche cruciali caratteristiche. Ogni attore del sistema sarà poi più o meno impegnato su ciascuno di essi.

Questo vissuto negativo rappresenta ciò che va comunque contrastato e per fare ciò esistono passi soggettivi e oggettivi. Dal lato dell'atteggiamento degli uffici di statistica va compreso che dobbiamo sostenere con forza ciò che appartiene alla funzione statistica, ma dobbiamo essere anche consapevoli che questa appartenenza può essere messa in discussione e va quindi conquistata giorno per giorno col consenso e l'autorevolezza. E' una richiesta di maturità, perché non ci sono più genitori che possono intervenire, ma al massimo alleati, che hanno tuttavia i loro obiettivi primari da perseguire.

Dal lato oggettivo, la nascita di sistemi informativi tematici fa parte proprio di quella razionalizzazione del Sistema necessaria per poter mantenere le caratteristiche delineate all'inizio. Oggi, una gran massa di informazioni disperse, senza sistemi di data mining, (per usare uno dei termini di questa

Conferenza), non riduce ma aumenta l'incertezza; esclude dal consumo, nel senso che solo i grandi esperti ci si possono orientare; richiede un gran lavoro di recupero per essere utilizzata efficacemente.

Questa evoluzione è, dunque, naturale e rappresenta in un certo senso quella che si cerca cooperativamente di accompagnare attraverso l'azione dei circoli di qualità.

3.4 Ufficio di statistica e forniture

Molti uffici di statistica, soprattutto centrali e, in generale, le amministrazioni cui appartengono, si sentono talvolta espropriati della loro base informativa quando il ministero o l'ente si alleggerisce di funzioni dirette attraverso acquisizione di forniture, che possono assumere la forma di contratti, convenzioni e concessioni.

Il tipico esempio è rappresentato dalla perdita informativa conseguente alla privatizzazione di settori ex monopolisti (telecomunicazioni, energia, ecc.). Normalmente in questi settori si ha un arricchimento informativo: le informazioni che oggi si acquisiscono sul traffico telefonico sono di gran lunga più ricche di quelle di un tempo, ma sono meno "pubbliche"; le informazioni sull'energia elettrica saranno più ricche e rimarranno pubbliche ma possedute da un nuovo gestore (GRTN); i contratti di servizio previsti per tutte le aziende in concessione nazionali e locali chiedono una quantità significativa di indicatori, ma essi sono sempre concepiti come informazione liberamente disponibile per i cittadini.

Negli anni '80 si propugnava la "copertura statistica" delle leggi. Oggi, per il pieno funzionamento del Sistan, per non disperdere preziose energie nel recupero ex post di informazioni perdute, bisogna rivendicare la copertura statistica delle forniture rilevanti, siano esse concessioni o contratti di servizio. Anche questa non rappresenta una novità, ma dopo una certa attenzione nei primi anni '90, non è stata più effettivamente perseguita.

Il tipo di specifiche informative necessarie e le modalità della fornitura dei dati possono essere progettati dall'Istat, ma forse anche da altri soggetti del Sistema direttamente coinvolti, e trasferite, come modello di riferimento, a tutti gli uffici di statistica che si trovano in condizione di volerne usufruire.

3.5 Uffici di statistica e personale politico

Tutto ciò che si è detto finora è pensato con l'obiettivo implicito di assicurare maggior potere negoziale agli uffici di statistica. Ma non è sufficiente. Il maggior disagio che traspare dall'interno degli uffici di statistica è la scarsa considerazione da parte del personale politico. Esiste solo un modo "tecnico" per intervenire: attuare una massiccia e intelligente formazione del personale politico.

L'impegno degli amministratori su numerosi fronti e sulle emergenze quotidiane, accanto ad una storica disattenzione verso la statistica e i dati, non giocano a nostro favore, anche se, continuamente verifichiamo delle eccezioni che tuttavia non fanno sistema. L'impegno che il Dipartimento del Sistan può assumere è quello di progettare azioni di comunicazione, formazione verso il personale politico, che abbiano positive ricadute sul clima in cui l'ufficio di statistica opera.

3.6 Prime conclusioni

La prima macrofunzione degli uffici di statistica è stata sviscerata in maniera più ampia perché ritengo sia la più delicata. Rappresenta quella in cui si confligge e si coopera con il resto dell'amministrazione di appartenenza. Ognuno fa le proprie scelte in casa propria, ma se il Sistema statistico nazionale si comporta come un corpo solidale, può emergere un forte aiuto, conoscendo meglio le situazioni similari, creando gruppi di scambio di esperienze, imitando e adattando pratiche migliori.

Siamo su un terreno che le teorie organizzative definiscono soft, ma non per questo meno importante. Tecnologie e legislazione ci sono alleate. Ambedue, di per sé, in assenza di pratiche contrastanti, spingono verso la condivisione e la libera circolazione delle informazioni raccolte e detenute.

Questa alleanza, fortemente ribadita in diversi interventi durante questa Conferenza, iniziando dalla relazione introduttiva del presidente Zuliani, è valida anche per l'altra macrofunzione che mi accingo ad illustrare brevemente: la produzione diretta di statistiche.

4. Uffici di statistica e attività di rilevazione/elaborazione

4.1 Rilevazioni/Elaborazioni dirette

Con questa seconda sfera di attività degli uffici, entriamo nel terreno centrale della loro funzione, quello di effettuare direttamente raccolte di dati, che si possono concretizzare in cinque principali tipologie di processo, secondo la figura 1.

Come si può evincere dall'albero di classificazione dei processi, esiste una diversità sostanziale dei processi di produzione delle statistiche. Vorrei fermarmi su uno solo di questi, sia per la sua maggiore problematicità, sia per il suo rapporto con le tecnologie dell'informazione, tema principale di questa Conferenza. Mi riferisco al processo indicato col numero 4: elaborazione su dati raccolti dalla/nella propria amministrazione.

Il ventaglio di situazioni è tra i più disparati. Da un lato abbiamo esempi di completa strutturazione degli archivi amministrativi e del loro utilizzo agevolato dalle tecnologie informatiche. Dall'altro abbiamo esempi di totale destrutturazione. Cito dalle conclusioni del Progetto di studio del sistema informativo della Regione Lazio, I.RE.LA.: "In ambiente statistico, un archivio è un insieme strutturato di informazioni disponibili in un dato momento su un particolare gruppo di soggetti (…). La funzione amministrativa tende invece a considerare l'archivio come l'insieme delle pratiche amministrative (…)".

Nella maggior parte dei casi incontrati nello studio sulla Regione Lazio, l'uso dell'informatica era pressoché nullo e l'approccio amministrativo nettamente prevalente. L'azione dell'informatica, che richiede almeno le stesse strutturazioni della statistica, potrebbe essere servente sia ad un processo che ad un altro, senza particolari difficoltà.

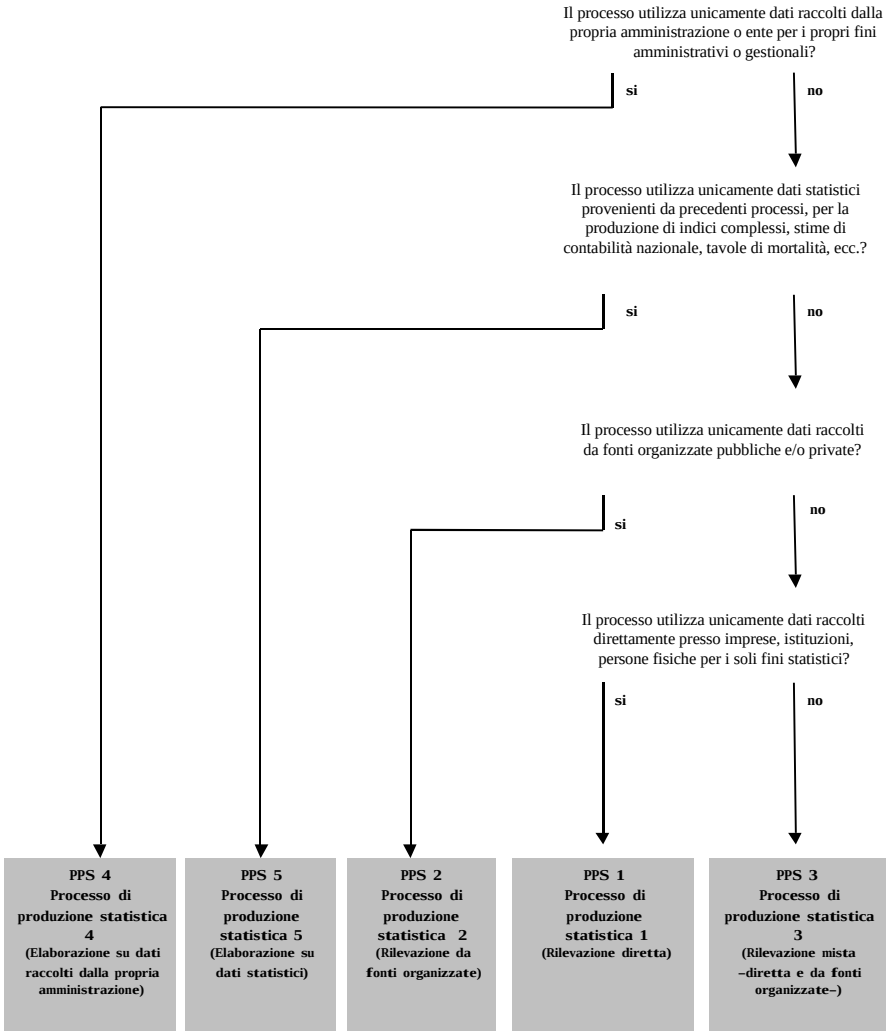
Si tratta di problematiche note da almeno quindici anni, ma ciò che sorprende è che sono ancora così presenti, in un contesto di non certo scarsa diffusione di basi dati informatizzate.

Il processo 4 della figura 1 raccoglie quindi situazioni molto diversificate. La check list di documentazione del processo, contenuta nel lavoro da cui è tratta la figura e preparata da un gruppo di studio misto, composto da rappresentati Istat e di enti Sistan, è il primo esempio di aiuto concreto per svolgere il ruolo diretto di reperimento di informazioni (sia come rilevazioni che come elaborazioni). Aiuterà nella consapevolezza del proprio lavoro, ad evidenziare i

punti critici e, soprattutto, a documentare il processo lavorativo per migliorarlo e "tramandarlo". Sarà di supporto anche per comprendere la fattibilità di certi programmi statistici: un ufficio di statistica di poche persone può accettare tra i suoi compiti l'estrazione e l'elaborazione di alcune tavole statistiche da archivi informatizzati (abbiamo visto in questi giorni, nel Salone dell'Informazione statistica, quanta strumentazione è disponibile e come il Sistema se ne possa appropriare); viceversa può non ritenere fattibile la preparazione delle stesse tavole a partire da pratiche amministrative.

Gli altri processi illustrati nella figura 1 possono portare a considerazioni più o meno simili alle precedenti.

Figura 1 - Albero delle decisioni per la classificazione dei processi



Questa check list (Manuale di documentazione) è forse il primo strumento sulla qualità che l'Istat rilascia, almeno in quella che possiamo definire la versione 1.0. L'Istat ha avuto una parte propulsiva, ma lo strumento è stato testato da diversi soggetti del Sistema. Ora si tratta di trasferirlo, preparare versioni successive e renderlo ampiamente utilizzato per poi procedere alla preparazione di altri: sui campioni, sulla correzione dei dati, sulla diffusione. Fin d'ora l'invito è ad attivarsi per collaborare, come portatori di pratiche, sperimentatori, applicatori.

4.2 Collaborazione a rilevazioni del Programma statistico nazionale

Il Psn ha superato le 1000 linee di lavoro e non ha ancora integrato quasi nulla della produzione territoriale. Le linee si raccolgono in alcune aree tematiche e sono sostenute da indicazioni strategiche che ne indirizzano l'azione. Teniamo presente che non è solo l'Istat ormai a chiedere collaborazione, ma tutto il Sistema che collabora, scambia, interscambia.

Queste linee di lavoro hanno cinquanta soggetti responsabili, ma si basano sul lavoro di migliaia di uffici di statistica. E' questa collaborazione, essenziale per la produzione delle statistiche, che fa il Sistema e il Psn.

Ponendoci sempre dal punto di vista dell'Ufficio di statistica che è chiamato a collaborare alle rilevazioni previste, senza esserne il responsabile primario, se siamo in una stagione di federalismo fiscale, non da meno siamo in una stagione di federalismo statistico. Federalismo fiscale vuol dire essenzialmente portare gli effetti vicino alle cause, pagare le tasse (cause) e vederne il riscontro (effetti). La rivendicazione statistica dell'ufficio, ma anche del singolo rispondente, è sempre più quella di voler vedere gli effetti della propria collaborazione.

Questa restituzione degli effetti, purtroppo, non può essere portata a livello così spinto come vorremmo. Bisognerebbe fare censimenti continui, essendo il censimento l'unico sistema che restituisce l'informazione fino alle cellule territoriali informative minime (le sezioni di censimento). La stagione dei censimenti è in corso e sappiamo tutti quanto sia complessa e faticosa. Ormai le tecniche statistiche e le tecnologie informatiche assicurano buone risposte, purché il materiale di base, cioè la raccolta diretta o tramite archivi sia buono. Questo è il terreno primario della cooperazione e della responsabilizzazione: fare in modo che il materiale di base sia buono e, quando è possibile, riorganizzarlo in modo da metterlo a disposizione delle comunità locali.

Qualcosa è stato fatto cercando di diffondere la pratica di Annuari e raccolte statistiche a livello comunale. In questa Conferenza sono stati presentati alcuni prototipi di Annuari e opuscoli divulgativi comunali. E' un terreno su cui

occorre insistere proprio per mantenere forte il legame tra collaborazione alle rilevazioni di dati e utilità delle statistiche.

Altro punto particolarmente delicato, ma anche grandemente fecondo è quello dell'integrazione tra Programma statistico nazionale e Programmi regionali, ovvero, in generale, l'integrazione con altri anelli informativi con caratteristiche specialistiche. Queste integrazioni possono e debbono portare ulteriori razionalizzazioni. Lo stato attuale è che un Programma statistico regionale cerca di delineare le informazioni e quindi le rilevazioni di interesse regionale; ma quante sono le informazioni di esclusivo interesse regionale? Le informazioni sulle materie decentrate alle Regioni sono di interesse regionale? Ovvero, proprio perché decentrate, sono di interesse nazionale o interregionale?

Riprendendo la metafora del federalismo fiscale, uno dei concetti fondamentali delle teorie in materia è quello del "traboccamento" (spillover) degli effetti. Le teorie cercano di definire le dimensioni territoriali più adatte a ridurre al minimo il traboccamento degli effetti, sia negativi che positivi, delle azioni del livello di governo corrispondente a quella misura territoriale. Ma, nell'informazione statistica, quando è che si può ritenere che una informazione di dettaglio comunale non sia più utile a livello regionale o nazionale, o che l'informazione su una regione sia inutile per la regione vicina? La statistica è fatta di confronti e, quindi, il traboccamento è sempre consistente.

Questo spinge ancora di più ad essere sistema e ad organizzare i programmi statistici in modo da tener conto, contemporaneamente, di responsabilità e di utilità. Se si accetta questa linea di ragionamento, la strada per l'integrazione dei Programmi statistici regionali, locali e funzionali nel Psn non è poi così complessa. In sostanza è il documento complessivo del Psn che va pensato diversamente, strutturato come un programma multiindice in cui, ad esempio, scorrendo le rilevazioni riguardanti una specifica regione, si trova quello che è di interesse (quasi) esclusivo della regione, quello che fa parte di una cooperazione interregionale per giungere ad indicatori comuni, quello che interessa il sistema Italia e che può avere diversi gradi di utilità a livello regionale. Infine, senz'altro, qualche linea di rilevazione darà un contributo alla statistica europea.

5. Alcune scelte di fondo e conclusioni

5.1 Priorità e razionalizzazioni

Nell'introduzione ho cercato di giustificare i motivi per cui la produzione statistica deve essere soggetta a ripetuta interrogazione sulla sua ragion d'essere, sul senso delle domande che vengono poste, sulle motivazioni conoscitive

alla base delle rilevazioni. Queste ultime (le motivazioni) non debbono essere per forza immediate e universali; possono anche essere documentali, intermedie e finalizzate ad ulteriori elaborazioni, coercitive o facoltative. Non possono tuttavia essere implicite o addirittura segrete. Sarebbe un paradosso per la statistica pubblica.

Più concretamente saremo costretti a partire dal prossimo anno a lavorare sulle priorità. "Le nuove esigenze statistiche dovranno essere fissate sulla base di effettive priorità e rilevanza politica, come pure sulla disponibilità delle fonti e la necessità di ridurre i carichi di risposta. Con l'emergere di nuove esigenze statistiche, la richiesta attuale dovrebbe essere riconsiderata per eliminare quelle aree di bassa priorità".

Sapere perché si fa qualcosa è l'unico modo per scegliere. Nel nostro caso, perché una rilevazione viene portata avanti. Come affrontare questo tema senza dover giungere a traumatiche e forse pericolose rinunce?

Specializzando gli organi del Sistema, la chiave di volta va cercata nel trasferimento progressivo di responsabilità significative verso il livello informativo in corrispondenza del quale va garantito il ritorno informativo maggiore, assicurando dal centro l'omogeneità dei metodi e delle tecniche. Tutto ciò si può fare solo se le ragioni di una rilevazione sono esplicite e si possono confrontare con quelle consimili. Si possono rivedere nel tempo, sono integrate a livello internazionale, si possono "giudicare" e consentono di formulare una scala di priorità. Nei termini indicati precedentemente, documentate.

Solo così il Sistema potrà allargare la propria area conoscitiva e toccare anche le tematiche sollevate in questa Conferenza, in particolare, misurare la cosiddetta società dell'informazione.

In sintesi, il cosa e il perché ci portano a integrare i Programmi statistici, ad esplicitare le motivazioni alla base delle rilevazioni, a rinforzare il lavoro dei circoli di qualità, a "tagliare" e "inserire" sulla base di valutazioni di priorità. Non perché questo non viene attualmente fatto: il turnover lordo di lavori nel Psn (entrati + usciti) si attesta sul 40%, cifra indicante un continuo lavorio, ma anche il sospetto di margini ampi di miglioramento della programmazione e della consapevolezza delle scelte.

5.2 Conclusioni

Non posso che concludere con l'invito della Commissione di Garanzia ad integrare il Psn con la previsione di un calendario dei rilasci dei dati e delle informazioni prodotte; l'Istat lo fa da qualche tempo ed è opportuno che questa pratica sia diffusa ad altri soggetti del Sistema. Questo potrebbe essere il naturale completamento della strumentazione operativa del Sistan. Inversamente

può sembrare che si fornisca solo il disturbo delle rilevazioni. Per la qualità che ormai comincia ad avere il Psn, è opportuno, inoltre, pensare già nel modo di presentarlo, a far emergere con più chiarezza i momenti diffusivi.

Il Dipartimento della segreteria centrale del Sistan ha elaborato un prototipo di sito/portale che consente di chiudere il cerchio tra programmazione di rilevazioni, contenute nel Psn, loro caratteristiche metodologiche, documentazione, presenza nel catalogo editoriale e collegamento con il sito, del soggetto responsabile delle rilevazioni. Questo prototipo potrà essere sviluppato solo se tutti i soggetti del Sistema garantiscono la costante manutenzione delle sue pagine. In sostanza, solo se diviene un portale multicontribuito.

Una delle critiche che ho sentito muovere al Sistan è quella di essere inconsistente, leggero, poco radicato. Sarà senz'altro in parte così; ma allora trasformiamo questo difetto in una opportunità. Se è leggero può correre veloce e, in poco tempo, radicarsi di più e meglio. La rete, adottata spesso come simbolo per il Sistema e rappresentata in qualche modo nel logo che abbiamo scelto, può essere solida e nello stesso tempo leggera, ariosa, non costrittiva. Ha un solo principio che la contraddistingue, quello della cooperazione; senza di questo, i nodi si sciolgono e la rete si indebolisce. Cooperare per costruirla mettendo a disposizione esperienze e conoscenze è facile e anche stimolante.

Lo dico perché pochi sistemi nazionali hanno questa ricca opportunità di colloquio, collaborazione, imitazione, gemellaggio, adattamento senza i vincoli costrittivi di essere uguali. Le ricche opportunità vanno sempre colte.

L'ATTIVITA' DELLA COMMISSIONE PER LA GARANZIA DELL'INFORMAZIONE STATISTICA: UN QUADRO DI SINTESI E SPUNTI DI RIFLESSIONE*

Ugo Trivellato

presidente della Commissione per la garanzia dell'informazione statistica

1. Introduzione

L'istituzione della Commissione per la garanzia dell'informazione statistica si colloca nel quadro dei cambiamenti nell'organizzazione della statistica ufficiale italiana, che sono sopravvenuti con la legge di riforma del 1989. L'idea-guida del legislatore, un'idea-guida che non mutua un "modello" consolidato in altri Paesi - al più trova qua e là analogie - , quindi in buona parte innovativa, è stata quella di prevedere, accanto agli enti cui è affidato il compito di produrre l'informazione statistica ufficiale, un organo autonomo, indipendente, con funzioni di vigilanza sull'insieme di tale produzione.

Tale vigilanza verte essenzialmente su tre domini:
- l'imparzialità, la completezza e la qualità delle informazioni statistiche prodotte dal Sistema statistico nazionale (Sistan);
- la conformità delle rilevazioni svolte dagli enti del Sistan, in primo luogo dall'Istat, alle direttive degli organismi internazionali e comunitari;

*Questa nota ha beneficiato delle discussioni e della riflessione collegiale in sede di Commissione per la garanzia dell'informazione statistica, e qua e là richiama pareri e posizioni della Commissione. L'insieme delle considerazioni prospettate, tuttavia, impegna soltanto l'autore.

- la tutela della riservatezza delle informazioni fornite a tali enti, rivisitata alla luce della recente normativa sulla privacy[1].

La Commissione – conviene precisarlo – ha il compito di esercitare una sorveglianza a carattere generale sull'attività svolta dagli enti del Sistan. Non indugio sull'argomento, perché è stato ben tratteggiato da Bernardo Colombo, nella relazione tenuta quattro anni or sono alla Terza Conferenza Nazionale di Statistica [2], alla conclusione del suo mandato di Presidente. Basterà ricordare qui che la Commissione, anche per i marcati tratti di novità che ne connotano la collocazione e le funzioni, "*ha dovuto dedicarsi ad una riflessione appro - fondita (…) per capire bene che cosa il legislatore le dava da fare, e come*". L'esito di tale riflessione, che si è venuto chiarendo ulteriormente a fronte di concrete questioni delle quali la Commissione è stata investita, è appunto una lettura della funzione di vigilanza in termini di "*sorveglianza a carattere infor - male, in contemporanea o successiva, dell'attività complessiva dell'ente con - trollato*", con un controllo ad referendum, il riferimento essendo innanzitutto il Presidente dell'Istat - per la sua posizione al vertice del Sistan oltre che dell'Istat stesso - e, se del caso, il Presidente del Consiglio dei Ministri.

2. I modi di operare della Commissione

Il momento più significativo dell'azione della Commissione è rappresentato dalla formulazione del parere sul Programma Statistico Nazionale (PSN). Esso consente, e insieme richiede, alla Commissione di vagliare l'adeguatezza delle realizzazioni e dei programmi dell'intero sistema delle statistiche ufficiali, sotto i profili, già menzionati, dell'imparzialità, della completezza, della qualità, dell'aderenza alle direttive dell'Unione Europea (UE). Nello stesso tempo, la Commissione si trova nella condizione di poter formulare osservazioni e suggerimenti sui programmi di attività dell'Istat e degli altri enti del Sistan. Il parere sul PSN finisce quindi per rappresentare la sede privilegiata anche per una più ampia funzione di riflessione e di stimolo in tema di sviluppo dell'informazione statistica ufficiale.

Il parere sul PSN, pur di primaria importanza, non esaurisce tuttavia l'azione della Commissione. Essa può predisporre note, raccomandazioni o rilievi in merito a singoli temi, che abbia giudicato meritevoli di una specifica sorveglianza vuoi per iniziativa propria vuoi a seguito di quesiti ricevuti.

La formulazione del parere sul PSN e questo diversificato insieme di

1. Vedi la legge n. 675/1996 e i successivi decreti legislativi di integrazione e correzione, fra i quali spiccano il decreto l.vo n. 135/1999 sul trattamento di dati sensibili da parte dei soggetti pubblici e il decreto l.vo n. 281/1999 sul trattamento di dati personali per finalità storiche, statistiche e di ricerca scientifica.
2. Vedi B. Colombo, "L'attività della Commissione per la garanzia dell'informazione statistica", Atti della Terza Conferenza Nazionale di Statistica. Roma, 26-27-28 novembre 1996, Roma, Sistan-Istat, 1999, pp. 167-176. I brevi passi riportati nel seguito del capoverso sono tratti da questa relazione.

interventi richiedono di poggiare su una continua attività di documentazione e di studio: da un lato per realizzare un monitoraggio corrente dello stato della statistica ufficiale; dall'altro per approfondire specifici settori o temi, che emergano con rilievo particolare o che manifestino aspetti problematici. Varie circostanze di indole organizzativa hanno a lungo condizionato i lavori della Commissione (e, a ben vedere, frappongono ancora oggi ostacoli a una sua piena capacità operativa). Ma miglioramenti ci sono stati, e apprezzabili: tra l'altro con il consolidamento, a fianco della segreteria amministrativa, di una segreteria tecnica. Ciò ha consentito di affrontare con una certa sistematicità i compiti istruttori, approntando vari strumenti di lavoro. Tra questi, due in particolare meritano qualche ragguaglio:

- lo svolgimento di audizioni su specifici argomenti, con previa predisposizione di quesiti di riferimento e con successive sintesi;
- la promozione di ricerche, anche con l'affidamento di incarichi ad esperti, mirate ad indagare lo stato dell'informazione statistica in specifici settori o la qualità di particolari fasi del processo di produzione dell'informazione stessa o le opportunità aperte per grandi tematiche.

Negli ultimi due anni, la Commissione ha tenuto audizioni (generalmente con dirigenti dell'Istat, in un caso con l'Autorità garante per la concorrenza ed il mercato) su diversi argomenti: informazione statistico-economica a livello territoriale; informazione statistica sul lavoro; procedure di destagionalizzazione di serie storiche economiche; progettazione dei censimenti del 2000-2001; adeguatezza delle informazioni disponibili per valutare le condizioni di concorrenza nei mercati; qualità delle operazioni sul campo e nuove tecniche di rilevazione *computer-assisted* in indagini sulle famiglie.

Affronterò il tema dei gruppi di ricerca, di quelli che hanno concluso l'attività in quest'ultimo biennio, quando accennerò alla diffusione dei risultati dell'azione della Commissione. Attualmente, sono attivi una dozzina di gruppi, che operano su un ampio ventaglio di temi: informazione statistico-economica territoriale; attuazione del SEC 95; confronti e possibilità di *arbitrage* dei dati sulla distribuzione commerciale e delle stime dei consumi familiari; misure di inflazione e sistema di monitoraggio prezzi; statistiche sugli investimenti stranieri diretti; informazioni statistiche sulla struttura e la dinamica delle piccole e medie imprese manifatturiere; completezza e qualità delle informazioni statistiche utilizzabili per la valutazione della spesa pensionistica; esperienze in tema di indagini Multiscopo e di *European Community Household Panel*, nella prospettiva di un'indagine panel sulle famiglie; rilevazioni delle migrazioni internazionali e sistema informativo sugli stranieri; indicatori per il monitoraggio e la valutazione dell'attività sanitaria; effetti della qualità del frame sui risultati di

indagini basate sul campionamento da liste anagrafiche.

Gli approfondimenti tramite i gruppi di ricerca e le audizioni, tra l'altro sovente realizzati in maniera coordinata, sono un supporto essenziale per l'azione di vigilanza. Essi, infatti, concorrono in maniera decisiva a fornire una solida base conoscitiva per le valutazioni, i pareri e le raccomandazioni della Commissione.

Confortata dal consolidamento nei modi di operare, recentemente la Commissione ha ritenuto di dare maggiore visibilità alla propria attività e agli atti prodotti [3]. Ciò ha richiesto la ricerca di un equilibrio fra due esigenze: il doveroso riserbo connesso alla funzione di sorveglianza e di controllo ad referendum; l'impegno a mettere a disposizione della generalità degli operatori e degli studiosi gli approfondimenti conoscitivi e gli orientamenti della Commissione. La scelta è stata nel senso di una ragionevole apertura, che fa perno su tre canali di diffusione.

- È stato dato ulteriore impulso alla collana di Rapporti di ricerca, che accoglie i rapporti conclusivi degli studi di maggior rilievo promossi dalla Commissione. Nell'arco dell'ultimo biennio ne sono stati prodotti una ventina[4]. I "Rapporti di ricerca" sono inviati correntemente a un ristretto gruppo di destinatari istituzionali. Su richiesta, sono inoltre messi a disposizione di studiosi[5].
- È stata avviata la predisposizione di un Bollettino annuale, rivolto a un pubblico più vasto, che si propone di dare conto sinteticamente dell'attività svolta. Il primo numero, una sorta di numero speciale, perché presenta un sommario dell'azione della Commissione nei suoi primi otto anni di vita – con un maggiore dettaglio l'attività dell'ultimo anno –, è stato appena pubblicato[6].
-Infine, da qualche mese è stato attivato un sito web della Commissione, nell'ambito del sito della Presidenza del Consiglio dei Ministri[7].

---

3. Della propria azione, la Commissione ha sempre dato conto in maniera circostanziata nel rapporto annuale, presentato al Parlamento come allegato alla relazione sull'attività dell'Istat.

4. Le aree tematiche toccate sono: la disciplina legislativa e l'organizzazione operativa della statistica, a livello comunitario e nazionale; la tutela della privacy in relazione all'accesso a dati individuali per fini di ricerca – con attenzione sia agli aspetti normativi che a quelli tecnici –; le attività preparatorie e di progettazione dei censimenti del 2000-2001; singoli settori o aspetti delle statistiche ufficiali (emarginazione sociale, giustizia, lavoro, rifiuti, disponibilità di statistiche di genere); tematiche metodologiche e tecniche, riferite prevalentemente alle indagini campionarie sulle famiglie (qualità delle operazioni sul campo, caratteristiche dei non rispondenti, procedure di correzione-imputazione) e relative inoltre alla destagionalizzazione.

5. In prospettiva, se ne prevede l'inserimento nel sito web della Commissione.

6. Il Bollettino è pubblicato dal Dipartimento per l'Informazione e l'Editoria della Presidenza del Consiglio dei Ministri.

7. L'indirizzo è: www.palazzochigi.it/commissione_statistica/index.html.

3. Pareri e altri contributi della Commissione: alcune evidenze salienti

3.1. Sull'impianto e l'organizzazione del Programma Statistico Nazionale

Ho già sottolineato la rilevanza che, nell'azione della Commissione, assume il parere sul PSN. Ma non è questa la sede per richiamarne i contenuti. Penso utile guardare, invece, a due questioni di indole generale: l'evoluzione della struttura e del processo di formazione-realizzazione del PSN da un lato; e dall'altro, il procedere della riflessione sulle linee strategiche del Programma. Esse si prestano per illustrare la funzione di stimolo svolta dalla Commissione e l'interazione che, nella sequenza di Programmi e di pareri, si è venuta sviluppando tra la Commissione e l'Istat.

La riflessione della Commissione è stata guidata da un obiettivo: fare del Programma uno strumento via via più efficace, accentuandone i tratti di sistematicità e coerenza. Le osservazioni e i suggerimenti avanzati sui vari PSN configurano un'azione di costante sollecitazione, che vede la Commissione accompagnare il miglioramento dell'impianto del Programma. Soprattutto i due ultimi Programmi, per i trienni 2000-2002 e 2001-2003, segnano tappe significative di tale percorso. L'ampliamento della tipologia dei lavori, l'arricchimento della documentazione sulle caratteristiche metodologiche e tecniche degli stessi, il rendiconto sui lavori presenti nel precedente PSN e non ripresentati nel Programma corrente, la progressiva realizzazione del Sistema informativo del Programma (e dell'organizzazione del Sistan) ne sono i tasselli più importanti.

I risultati sono soddisfacenti soprattutto sul fronte dello sviluppo temporale delle attività previste dai successivi Programmi. Si dispone ora di un buon quadro conoscitivo del processo di prosecuzione-rettifica-innovazione delle attività, essenziale per monitore il PSN nella sua dimensione dinamica, il che è cruciale per un programma "scorrevole": con orizzonte triennale ma riformulato ogni anno.

Miglioramenti si registrano anche nell'integrazione strutturale, "trasversale" del PSN. Lo snodo critico è, qui, nel raccordo fra le linee strategiche e gli obiettivi generali, enunciati per l'intero sistema, e i progetti e i singoli lavori per settore di interesse, illustrati successivamente in maniera analitica distinguendo l'apporto dei vari soggetti del Sistan. Un importante contributo è venuto dai circoli di qualità, i quali vedono le varie componenti del Sistan e altri utenti dell'informazione statistica impegnati congiuntamente nella definizione delle esigenze informative e nell'elaborazione delle linee programmatiche. Ma su questo fronte resta ancora parecchia strada da fare. La traduzione operativa delle linee strategiche del PSN in un integrato quadro d'insieme di obiettivi, progetti e lavori non è ancora riconoscibile in maniera diffusa, omogenea. Emerge inoltre evidente l'esigenza che siano accentuate le caratteristiche di "trasversa-

lità" e l'integrazione di numerosi progetti, che coinvolgono lavori di settori ed enti diversi.

Ciò chiama in causa il tema di coordinamento tra i diversi soggetti del Sistan. Ai vantaggi forniti dalla positiva interazione fra attività dell'Istat e delle amministrazioni del settore, testimoniati da realizzazioni innovative in alcune aree; (si ricordano ad esempio, le esperienze di collaborazione dell'Istat con il Ministero dell'Università e della Ricerca Scientifica e con l'Istituto per il Commercio con l'Estero)[8], si contrappone in altri comparti una situazione di stallo. Luci e ombre, dunque. Il superamento di queste carenze richiede un impegno di lunga lena, che si dilata ad aspetti istituzionali e organizzativi e alla necessità di maggiori risorse: questioni sulle quali tornerò brevemente alla fine. Vorrei segnalare qui due preoccupazioni più circoscritte, direttamente collegate al mandato di vigilanza sulla qualità e l'imparzialità dell'informazione statistica, sulle quali la Commissione ha posto insistentemente l'accento.

- Da un lato, ha raccomandato di potenziare il ruolo di indirizzo, di guida e sorveglianza metodologica e di assistenza tecnica dell'Istat nei confronti della produzione statistica degli altri soggetti del Sistan. Il suggerimento ha trovato positivi riscontri nel PSN 2001-2003, che pre vede specifici progetti in questa direzione.
- Dall'altro lato ha sollecitato la predisposizione da parte di tutti i soggetti del Sistan di un calendario per la diffusione dei risultati delle rilevazioni/elaborazioni di competenza. Da parecchi anni lo fa l'Istat, e con lodevole rigore. Così non è per altri soggetti del Sistan, con la conseguenza non dirò di balletti di cifre, ma certo di uscite disparate, sovente intempestive. L'accavallarsi delle stime delle variazioni dei prezzi al consumo delle città capoluogo di regione, che precede di qualche giorno la stima nazionale, ne è l'esempio forse più vistoso: la pratica, pure legittima [9], è palesemente inopportuna, perché crea confusione e nervosismi, senza apportare vantaggio alcuno. L'esigenza di porre rimedio a questo stato di cose si impone oggi con particolare forza, in concomitanza col passaggio della gestione di importanti statistiche di settore dall'Istat agli uffici di statistica dei pertinenti Ministeri.

8. La convenzione fra Istat e Ministero dell'Università e della Ricerca Scientifica ha consentito di valorizzare le capacità di rilevazione statistica dell'amministrazione e di accentuare il ruolo dell'Istat in chiave di progettazione e assistenza tecnica e di svolgimento di indagini sulla transizione fra università e lavoro. L'esito della collaborazione fra Istat e Istituto per il Commercio con l'estero è il nuovo annuario statistico del settore.

9. La normativa che la consente risale al 1927, e anche sotto numerosi altri profili merita una revisione.

3.2. Sulle linee strategiche del Programma Statistico Nazionale

Venendo agli aspetti di merito, mi fermo su alcune delle linee strategiche degli ultimi Programmi, alle quali la Commissione ha dedicato notevole attenzione. Preliminarmente, debbo tuttavia richiamare, sia pure per rapidi tratti, le trasformazioni del contesto, entro le quali si colloca la riflessione sugli obiettivi della statistica ufficiale. Due sono i termini di riferimento essenziali.

- Uno è costituito dall'innovazione tecnologica, innanzitutto (ma non solo) in campo informatico, e dai profondi riflessi che essa ha sull'intero processo di produzione dell'informazione statistica - dalla raccolta, all'archiviazione, al trattamento, alla diffusione - . È il tema di questa Conferenza, e non vale certo la pena che vi indugi.
- L'altro riferimento è rappresentato dai cambiamenti nella domanda di informazione statistica, che vengono direttamente dagli operatori pubblici. In estrema sintesi: (i) procede con vigore l'armonizzazione della funzione statistica a livello dell'UE, che ormai evolve decisamente verso un vero e proprio sistema statistico comunitario; (ii) procede il trasferimento di funzioni a Regioni ed enti locali, che investe naturalmente anche i fabbisogni informativi; (iii) sul fronte del processo di riforma della pubblica amministrazione, si consolida l'attenzione all'attivazione di sistemi informativo-statistici di supporto alla programmazione, al controllo interno, al monitoraggio, alla valutazione delle politiche a livello centrale e locale.

In questo quadro, si pone con particolare forza l'esigenza che la statistica ufficiale accentui i tratti di "sistema": capace ad un tempo di rispondere a un corpo di domande diversificato - quanto a interrogativi ed a soggetti che li pongono - e di mantenere caratteristiche di coerenza e di integrazione. Questa esigenza è riassunta efficacemente nell'ultimo PSN, nella linea strategica della "costruzione di sistemi informativi integrati". Essa già annovera prime significative realizzazioni, tra le quali il sistema informativo delle statistiche sulle imprese e le istituzioni. La Commissione ha espresso apprezzamento per tale orientamento. Ha inoltre formulato suggerimenti mirati a rafforzare l'impegno in questa direzione, in particolare nel campo delle statistiche sociali [10].

10. Le indicazioni sono del seguente tenore: " Gli obiettivi del PSN [nell'ambito delle statistiche sociali]… riguardano propriamente sub-sistemi, sia pure di notevole importanza (quello su popolazione e salute, quello sulla sorveglianza integrata sui parti e gli esiti del concepimento, quello incentrato sulle indagini Multiscopo). Un allargamento della prospettiva, con l'estensione del 'frame' all'insieme delle informazioni su popolazione-famiglie-partecipazione al lavoro-redditi-consumi-condizioni di vita, è certo un'impresa assai impegnativa e di medio-lungo periodo. Ad essa è tuttavia indispensabile cominciare a porre mano, per disporre di un quadro informativo integrato dei fenomeni e processi sociali." (Commissione per la Garanzia dell'informazione statistica, "Parere sul programma Statistico Nazionale 2001-2003", Roma, 7 luglio 2000, p. 4).

Anche se il rischio di ribadire l'ovvio è appena dietro l'angolo, v'è da aggiungere che un sistema informativo statistico è la realizzazione più lontana che si possa immaginare da una mera banca dati, o da un insieme di banche dati interconnesse. L'informatica è uno strumento indispensabile; ed è uno strumento "intelligente", quindi da utilizzare al meglio. Ma resta uno strumento. I grandi avanzamenti tecnologico-informatici nel data capture, nell'organizzazione delle informazioni e nel loro trattamento aprono sì grandi potenzialità, ma non sono di per sé risolutivi per la bontà dell'informazione statistica. E per la statistica ufficiale in primo luogo, la quale deve obbedire a stringenti canoni di imparzialità e di qualità, non vi sono convincenti sostituti all'attenzione alla complessità del processo di misura, al rigore di definizioni e classificazioni, alla cura nella raccolta dei dati - vuoi "filtrando" quelli di origine amministrativa vuoi acquisendoli tramite surveys - , ad una organizzazione meditata del sistema informativo [11], all'integrazione nel sistema stesso di un adeguato corpo di metadati.

Agli inizi degli anni '70, quando facevano il loro ingresso nella statistica ufficiale i primi, agli occhi di oggi rudimentali, sistemi informativi statistici, contro i rischi di scomposte euforie conobbe una certa fortuna, forse anche perché condensato in un acronimo incisivo, un caveat: gigo, garbage in - garbage out. Da allora molta acqua è passata sotto i ponti, e oggi l'espressione suona certamente impropria, Ma un nocciolo di verità lo mantiene, ed è un utile monito per lo statistico ufficiale (e, a ben vedere, *tout court* per l'analista).

Assume perciò un valore particolare il fatto che nel PSN l'obiettivo della costruzione di sistemi informativi statistici sia affiancato da quello del "miglioramento della qualità dei prodotti e dei processi di produzione". Questo indirizzo è pienamente condiviso dalla Commissione. In particolare, essa ha dato una valutazione positiva dell'impegno dell'Istat nella direzione di metodologie generalizzate per il controllo e l'analisi della qualità. E un analogo giudizio la Commissione ha espresso per la linea di lavoro relativa all'automazione nell'acquisizione e nella codifica dei dati. Aggiungendovi tuttavia riflessioni, le quali mettono in luce, accanto ai vantaggi, gli ostacoli che si accompagnano alle innovazioni tecnologico-informatiche nelle modalità di acquisizione dei dati, ad esempio da *surveys*.

Prendiamo un caso emblematico: la tecnica CATI, cioè l'intervista telefonica assistita dal computer. Ad essa si è dato, e giustamente, notevole impulso, perché consente di disegnare il questionario in modo da contenere gli errori di risposta e permette inoltre di somministrarlo con un maggiore controllo della rete di rilevazione. Tradizionalmente, il tallone di Achille per indagini

11 Tale organizzazione non può prescindere, tra le altre cose, dalla combinazione di quadri concettuali - reti di relazioni e schemi contabili in cui esse si traducono - e della sottostante piattaforma di microdati.

CATI era rappresentato dal fatto che solo una parte – e una parte selezionata – della popolazione disponeva del telefono, il che induceva distorsioni nelle stime. Era opinione comune, tuttavia, che la progressiva diffusione del telefono avrebbe risolto la questione. Ora, la diffusione del telefono c'è stata. Ma lo svolgimento di indagini CATI si confronta con problemi, indotti dallo sviluppo della telefonia mobile e dalla molteplicità dei gestori, largamente inattesi: la crescente debolezza del frame costituito dagli elenchi telefonici e, per un altro verso, la difficoltà di ricorrere a tecniche di Random Digit Dial. Serve quindi una riflessione sull'argomento, a mente fresca perché i termini della questione sono largamente nuovi, la quale consenta di ricalibrare le strategie di rilevazione dell'Istituto.

Tra le iniziative per il miglioramento della qualità di prodotti e processi, nel PSN 2001-2003 trova poi rilievo l'impegno dell'Istat di fornire supporto metodologico agli altri soggetti del Sistan. Esso si concretizza in tre studi progettuali: sugli strumenti per la diffusione, presso tali soggetti, delle metodologie di indagine; sul software per la progettazione di campioni; sul database per l'estensione al Sistan Sistema informativo di documentazione delle indagini. Non sorprendentemente, la Commissione ha valutato in maniera assai positiva queste iniziative, anche perché rispondono a suggerimenti formulati in occasione di precedenti pareri. Vi ha aggiunto, tuttavia, una sollecitazione: che siano accompagnate da programmi mirati a stabilire standard su definizioni, nomenclature e classificazioni, sui tempi di diffusione dei risultati delle indagini, sulla documentazione che deve affiancarli.

3.3. Altri interventi

Nell'ultimo biennio la Commissione è intervenuta con pareri a fronte di specifiche richieste, con note e raccomandazioni in diverse circostanze, avendo come interlocutore il Presidente dell'Istat, in un caso il Segretariato Generale della Presidenza del Consiglio e in un paio di occasioni il Ministro per la Funzione Pubblica, destinatario della delega del Presidente del Consiglio in materia di statistica.

In generale, le note al Presidente dell'Istat si sono collocate a valle degli approfondimenti condotti con audizioni e gruppi di ricerca. Esse hanno riguardato le statistiche del lavoro, le procedure di destagionalizzazione di serie storiche economiche, la nuova contabilità SEC 95 - settori contraddistinti tutti da

marcate innovazioni nei processi e nei prodotti - [12], e soprattutto la preparazione dei censimenti.

Indugio brevemente sulle indicazioni formulate in materia di censimenti per un duplice motivo: per l'importanza che essi hanno per conoscenze dettagliate sulle caratteristiche strutturali della popolazione e delle attività economiche; perché la riflessione della Commissione ha riguardato sia gli aspetti progettuali e tecnici sia le risorse disponibili, le quali, palesemente, condizionano la completezza e la qualità dei risultati. Larga parte delle valutazioni nel merito del programma delle rilevazioni censuarie sono confluite, sia pure in forma sintetica, nel parere sul PSN 2001-2003, al quale dunque rimando. Sui margini di incertezza in merito ad un adeguato finanziamento dei censimenti, la Commissione si è soffermata già nei primi mesi di quest'anno, rilevando con preoccupazione la non ancora compiuta definizione della copertura finanziaria, testimoniata da uno scarto apprezzabile fra il fabbisogno quantificato dall'Istat e lo stanziamento dello Stato. Conseguentemente, ha segnalato la situazione al Ministro per la Funzione Pubblica. Da allora passi in avanti sono stati compiuti, ma restano ancora divergenze non trascurabili. L'auspicio è che sia assicurata la copertura del fabbisogno necessario per le operazioni censuarie progettate, e che siano così evitate dolorose rinunce - una sensibile contrazione del questionario del censimento della popolazione e delle abitazioni e una pesante riduzione delle indagini di qualità -. Delle conseguenze di queste rinunce, quando in un prossimo futuro risultassero indisponibili informazioni basilari, tra l'altro, per la buona amministrazione di comuni, regioni e dell'intero paese, resterebbe per certo la possibilità di rammaricarsi, difficilmente quella di porvi rimedio.

Un aggiornamento, e qualche commento, merita, infine, l'attività della Commissione in tema di tutela della riservatezza dei dati personali forniti agli enti del Sistan. Mi è capitato di sentirmi dire che in questi ultimi anni la Commissione ha mutato il proprio atteggiamento, passando da una posizione ispirata ad una forte preoccupazione per il rischio di identificazione, quindi restrittiva sulla pubblicazione di dati disaggregati e sul rilascio dei cosiddetti

12. In generale, i progetti e le realizzazioni dell'Istat sono stati valutati in maniera positiva. Quanto ai singoli settori, in materia di statistiche del lavoro la Commissione ha suggerito che l'introduzione della nuova indagine sulle forze di lavoro avvenga con la sovrapposizione di un anno con l'attuale, per consentire un raccordo esteso alla componente stagionale. In tema di destagionalizzazione, la Commissione ha sollecitato approfondimenti in merito alla scelta di un approccio diretto per gli indici della produzione industriale (mentre per gli altri settori è adottato l'approccio indiretto); inoltre, ha suggerito di allargare il numero di serie trattate, sulla falsariga di quanto raccomandato dal regolamento comunitario sulle statistiche congiunturali. Sul sistema SEC95, la Commissione ha espresso particolare apprezzamento per il rispetto degli obblighi comunitari in materia di diffusione dei nuovi conti e per la documentazione prodotta; ha poi suggerito che, nel quadro dell'attività di revisione e potenziamento dei conti in corso, siano approfondite questioni attinenti ai conti dei settori istituzionali (in particolare a quelli della Pubblica Amministrazione) e a una maggiore tempestività e dettaglio dei conti territoriali.

file standard[13], a un atteggiamento più liberale, attento in particolare alle esigenze della ricerca scientifica. Si tratta di un'impressione sbagliata. Quel che è cambiato è il contesto, a seguito dell'approvazione della legge sulla tutela della privacy. Con due conseguenze di rilievo: da un lato le funzioni di puntuale controllo e di sanzione (che comunque, nell'interpretazione data dei compiti di vigilanza, non erano mai state proprie della Commissione) fanno chiaramente capo al Garante per la tutela dei dati personali; dall'altro - e quel che più conta - , la definizione delle regole per l'accesso a dati personali per finalità statistiche e di ricerca diventa in parte questione de iure condendo. Il legislatore, infatti, affianca alla legge generale sulla tutela della privacy una legge di delega al Governo a integrare e correggere la normativa appena approvata, tra l'altro specificamente per "*il trattamento di dati personali utilizzati a fini (…) di ricerca e di statistica*"[14].

La Commissione si è trovata così ad operare su due distinti fronti. Per un verso, ha continuato a vigilare sui criteri seguiti dall'Istat per il rilascio dei file standard e ha seguito con attenzione la realizzazione, in via sperimentale, di un sito "sicuro" per il trattamento di dati personali. Per un altro verso, è stata coinvolta, informalmente ma fattivamente, nelle consultazioni per la predisposizione della normativa per il trattamento di dati personali utilizzati a fini di statistica e di ricerca: dapprima l'apposito decreto legislativo; successivamente l'elaborazione dei codici deontologici, previsti dal decreto stesso, che affida loro un'ampia funzione nello specificare la disciplina [15].

L'azione della Commissione ha continuato a fare perno sull'obiettivo di tutelare contemporaneamente la riservatezza e il diritto all'informazione statistica, bilanciando in maniera opportuna i due "valori" tramite un insieme coordinato di norme, strumenti tecnici, regole deontologiche [16]. Muovendo da questo punto fermo, si è mirato a mettere fuoco i desiderabili contenuti dei codici deontologici, alla luce delle "buone pratiche" sviluppate in vari Paesi [17] e degli avanzamenti nella tecnologia del trattamento delle informazioni (avanzamenti che - si noti - non agiscono solo nel senso di aumentare i rischi di identificabi-

13. Così sono designate le collezioni campionarie di dati elementari resi anonimi messe correntemente a disposizione dall'Istat.

14. Art.1, comma 1, lettera a) della legge n. 676/1996.

15. Il decreto l.vo n. 281/1999 prevede poi che, a conclusione di questo percorso, la Commissione sia "sentita ai fini della sottoscrizione del codice di deontologia (…) relativo al trattamento di dati personali nell'ambito del Sistan" (art. 12, comma 6).

16. A seguito della legislazione sulla privacy di dicembre 1996, la Commissione ha precisato i propri orientamenti in una nota ad uso interno del febbraio 1997. Le opzioni di fondo si ritrovano nella relazione di Luigi Biggeri, allora Presidente, alla Quarta Conferenza Nazionale di Statistica: vedi L. Biggeri "Diritto alla privacy e diritto all'informazione statistica", Atti della Quarta Conferenza Nazionale di Statistica. Roma, 11-12-13 novembre 1998, Roma, Sistan-Istat, 1999, Tomo 1, pp. 259-279.

17. Sull'argomento rimando a U. Trivellato, "Data access versus privacy: an analytical user's perspective", Statistica, 60 (3), 2000, in corso di pubblicazione.

lità, ma offrono anche crescenti possibilità di elaborazione “protetta” di dati personali e di prevenzione e c ontrollo dei tentativi illeciti di identificazione). La Commissione ha inoltre operato per favorire il confronto fra operatori statistici, comunità scientifica e Garante e, in collaborazione con la Facoltà di Scienze Statistiche dell’Università di Roma “La Sapienza” e con il Consiglio italiano per le Scienze Sociali, nell’aprile scorso ha organizzato un Seminario su “*Privacy*, statistica e ricerca: i codici deontologici”.

Quali gli esiti di questa azione? Nel marzo di quest’anno, sono state presentate al Garante due proposte di codici deontologici per il trattamento di dati personali, rispettivamente nell’ambito del Sistan e per scopi statistici e di ricerca scientifica: due proposte elaborate in maniera coordinata, che condividono i princìpi ispiratori e parecchie disposizioni. L’ulteriore confronto promosso dal Garante ha permesso di affinarle e di giungere alle proposte finali. È ora desiderabile che il Garante concluda sollecitamente il processo di sottoscrizione dei due codici e ne curi la pubblicazione nella Gazzetta Ufficiale. L’applicazione del corpo sostanziale della normativa per fini statistici e di ricerca, con le maggiori opportunità di trattamento dei dati individuali che consente, è infatti condizionata al fatto che i codici divengano efficaci.

4. Osservazioni conclusive

Avviandomi a concludere, prospetto alcune considerazioni su due temi di indole più generale. Sono temi che attengono ad assetti istituzionali e organizzativi, ed esulano quindi dall’ambito diretto della vigilanza della Commissione. Ma non si tratta di un’incursione fuori campo. Va da sé, infatti, che carenze sul fronte dell’organizzazione possono condizionare, anche pesantemente, la qualità di indagini, elaborazioni e prodotti. Muovendo da riscontri insoddisfacenti - o da motivate preoccupazioni - sui risultati, non si può non guardare anche alla dimensione organizzativa. Certo, non per proporre puntuali soluzioni. Ma sicuramente per segnalare esigenze funzionali, che emergano nitidamente dall’azione di sorveglianza affidata alla Commissione e che richiedano di trovare coerenti risposte anche sul terreno organizzativo.

Già ho segnalato come l’armonizzazione della funzione statistica a livello dell’UE proceda a grandi passi. Ciò risalta chiaramente dalla dilatazione dei regolamenti sulle statistiche comunitarie. L’elemento che mi preme mettere in luce è, tuttavia un altro, complementare. Poggiando anche sulle nuove disposizioni del trattato di Amsterdam, sono stati avviati, in interazione fra Consiglio, Comitati di Ministri di settore e Commissione dell’UE, processi per sostenere e monitorare politiche.

Il nucleo forte è rappresentato dall’azione dell’Ecofin, che, tra l’altro,

sovrintende alla formulazione dei fabbisogni informativi per l'Unione Monetaria Europea. Essa si è recentemente tradotta nell'"Action Plan on EMU Statistical Requirements", il quale individua azioni per migliorare la tempestività e la copertura delle statistiche congiunturali nell'UE, e conseguentemente specifica stringenti priorità per ciascun Paese.

Analoghe iniziative sono state avviate anche per settori dove la competenza diretta dell'Unione è minore, e si configura prevalentemente in termini di promozione della cooperazione fra gli Stati membri e di azione integrativa. È questo il caso del "processo di Lussemburgo" del 1997, relativo all'occupazione, e del "processo di Lisbona" di quest'anno, che riguarda per un verso la coesione e l'inclusione sociale e per un altro verso l'innovazione e la ricerca.

Il tratto che accomuna tutte queste iniziative è l'avvio di processi di "coordinamento aperto", basati sui seguenti elementi:

- identificazione di indicatori condivisi;
- elaborazione di linee guida comuni ed obiettivi quantificabili;
- presentazione di piani nazionali annuali;
- monitoraggio e benchmarking a livello comunitario.

In buona sostanza, siamo di fronte a processi "ad alta intensità di indicatori". Essi hanno conseguenze di grande rilievo per i sistemi statistici nazionali, conseguenze che sarebbe pericoloso venissero sottovalutate.

È necessario che l'Italia si attrezzi per partecipare in modo efficace a queste nuove forme di coordinamento delle politiche, anche per quanto riguarda il supporto di informazioni statistiche che deve alimentarle. Le implicazioni per il nostro sistema statistico sono parecchie, e sostanziose. Esse vanno dal raccordo ex ante fra Ministeri competenti, rispettivi uffici di statistica (in quanto esistano, e per davvero!) ed Istat, all'attivazione di strutture tecniche capaci di svolgere le funzioni conoscitive e valutative, alla messa a disposizione dei necessari finanziamenti.

L'argomento delle ricadute organizzative degli impegni comunitari si innesta poi nella più generale prospettiva di una revisione della normativa sul Sistan. Una prospettiva che, a dieci anni di distanza dall'istituzione del Sistan, è sollecitata dall'esperienza: dai difetti mostrati dall'attuale assetto e dalle debolezze nella sua attuazione [18] e che ha come rilevante elemento di sfondo il decentramento dei poteri verso la scala regionale e locale. In proposito, vorrei semplicemente fare presente una preoccupazione. Occorre evitare una prospettiva angusta, incentrata su un disegno di riforma del sistema statistico tutto giocato sulla dimensione "verticale" - lo Stato, le Regioni, le Province, i Comuni - polarizzato su una logica patrimoniale nell'acquisizione e nell'uso dei dati da

18. Tra le ragioni di debolezza (o tout court di inesistenza) degli uffici di statistica di parecchie amministrazioni, vi è anche il fatto che la riforma fu prevista, come si suole dire, "a costo zero" (art. 26, comma 3 del decreto l.vo n. 322/1989).

parte dei diversi soggetti.

Quando si sottolinea l'esigenza di disporre di un'informazione statistica maggiormente rispondente alle domande locali, a ben vedere si chiamano in causa due spazi:

- quello delle partizioni "regionali", intese in senso largo, nelle quali si articola l'organizzazione amministrativa, dotate di compiti e di poteri: le regioni, le province, le città;
- quello dei riferimenti funzionali, definiti da aree di interazione che variano in relazione a specifici processi sociali e a specifiche politiche: i mercati del lavoro locali, i distretti industriali, le aree di crisi.

Ora, è fuor di dubbio che la domanda riferita alle partizioni amministrative ha un notevole rilievo. Ma le possibilità di articolazione territoriale del sistema statistico non possono, e non debbono, essere disegnate soltanto lungo quest'asse. Non c'è, infatti, un mapping uno a uno fra spazi (e questioni) funzionali e partizioni amministrative, tipicamente regionali. E, d'altro canto, le stesse partizioni regionali si connotano sempre più come ambiti aperti, interdipendenti.

Di conseguenza, il sistema statistico deve esse re per quanto possibile flessibile, in modo da poter rispondere alle molteplici necessità conoscitive, riferite a scale territoriali diverse e variabili. Ciò porta in primo piano, anche per la riforma del Sistan, esigenze di coerenza, integrazione, modularità. D'altra parte - e i lavori di questa Conferenza lo hanno messo in luce con netta evidenza - , l'innovazione tecnologico-informatica permette ormai di distribuire l'informazione statistica in maniera facile, rapida, duttile; diminuisce, quindi, drasticamente l'ancoraggio dell'informazione stessa a una localizzazione fisica, così come il privilegio di un soggetto nell'utilizzarla.

Oltre che sull'architettura più o meno decentrata dell'assetto organizzativo, varrà dunque la pena di avere ben presenti anche altre questioni:

- come attivare adeguati canali di comunicazione con i principali soggetti nella fase di espressione della domanda, delle esigenze conoscitive, e quindi della definizione di obiettivi e programmi;
- come far crescere competenze e capacità professionali, come utilizzarle entro una trama organizzativa razionale, come rafforzare l'autonomia della funzione statistica ufficiale, perché possa produrre in misura e qualità adeguate un essenziale bene pubblico.

LA NUOVA ORGANIZZAZIONE DELL'ISTAT, UNA RISORSA PER IL SISTEMA STATISTICO NAZIONALE

Giuseppe Perrone

direttore generale dell'Istituto nazionale di statistica

1. La nuova organizzazione dell'Istat

Il modello organizzativo dell'Istat fino a qualche mese fa non separava in modo netto la funzione di indirizzo e controllo dalla funzione gestionale; tuttavia, attraverso gli istituti della delega e del decentramento, ampiamente utilizzati, le riforme degli ultimi anni nella Pubblica Amministrazione erano state recepite nella sostanza.

L'Istituto è stato tra i primi enti ad adeguare, fin dal 1998, il proprio bilancio ai principi contenuti nella legge 94/97 che prevede l'introduzione delle unità previsionali di base; il bilancio così strutturato è finalizzato, oltre che ad un'idonea allocazione delle risorse, ad individuare esigenze e di conseguenza titolarità e responsabilità.

La riforma organizzativa dell'Istituto è avvenuta in piena armonia con le norme esistenti e il nuovo regolamento ha tra l'altro definito la separazione tra organi di governo e uffici dirigenziali operativi, ha fissato compiti e competenze di entrambi, delineando un quadro di equilibrio in cui il decentramento si coniuga con il coordinamento e l'indirizzo generale.

La nuova organizzazione dell'Istat consentirà anche di servire meglio il Sistema statistico nazionale. Questo intento è già espresso in un atto organizza-

tivo generale – Linee fondamentali di organizzazione – allorché si è enunciata l'attività di promozione del Sistan.

Le **aree di produzione statistica** hanno subìto una importante razionalizzazione con la creazione di sei direzioni centrali incardinate nei due dipartimenti economico e sociale: interlocutori riconoscibili per un complesso di materie e una visione unitaria e non dispersa su tematiche affini.

In secondo luogo, il **dipartimento dell'integrazione degli standard tecnici**, che ha riunito diverse funzioni, dagli studi metodologici alla diffusione, dall'informatica alla contabilità nazionale, è chiamato a collaborare con il **dipartimento della Segreteria del Sistan**, per elaborare metodi e standard da mettere a disposizione dei soggetti del Sistema.

Infine, la Segreteria centrale del Sistan è divenuta dipartimento ed è stata rafforzata con la costituzione di una direzione centrale dedicata al coordinamento degli uffici regionali dell'Istituto; tra i compiti di questi ultimi entra, con grande enfasi, la collaborazione allo sviluppo locale del Sistema.

2. L'Istat a sostegno del Sistema

Con la nuova organizzazione, l'Istat ha rafforzato il suo ruolo di soggetto funzionale del Sistema precostituendo le condizioni per affrontare la sfida dell'**integrazione** e la sfida della **qualità**.

Da un punto di vista organizzativo, l'integrazione comporta la presa di coscienza circa la crescita progressiva delle autonomie e con l'esigenza di valorizzare la loro espressione e, nello stesso tempo, predisporre gli strumenti che consentano al Sistema di crescere in termini equilibrati.

Occorre impegnarsi a far crescere una cultura della "facilitazione organizzativa" e per quanto riguarda il conferimento di risorse, il passaggio da assetti rigidi coerenti con una logica di responsabilità centralizzata ad assetti flessibili che privilegino il partenariato e la cooperazione.

"La sfida della qualità" si connette con quella dell'integrazione. Quanto più si riuscirà a costruire circuiti orientati a elaborare, condividere e rispettare gli standard, tanto più si riuscirà a valorizzare iniziative che mettano in condizione i diversi partners di offrire il meglio delle proprie competenze in un contesto di rete.

Al di là degli aspetti organizzativi, integrazione e qualità rappresentano una **sfida sotto il profilo dei valori**.

Difficilmente si realizza un'efficace integrazione senza consapevolezza della logica di sistema, senza visione di sistema. Difficilmente si raggiunge qualità diffusa se non esiste identità di scopo fra gli attori delle diverse fasi procedurali e organizzative, in qualsiasi esperienza produttiva.

3. Le risorse

Abbiamo a disposizione strumenti privilegiati per sviluppare la cultura dell'integrazione e la crescita della qualità.

Alcuni di essi derivano dal contesto normativo, altri da una cultura dell'informazione statistica che si è rinnovata in maniera significativa negli ultimi anni, altri ancora dall'aver piegato in opportunità alcune sfide particolarmente onerose che il sistema si è trovato e si trova ad affrontare. Altri, infine, scaturiscono dall'innovazione tecnologica.

La normativa ci indirizza verso lo sviluppo di iniziative di partenariato e in questo senso va ricordato in particolare il decreto legislativo 419 del 1999, con il quale si è dato spazio all'individuazione di forme di cooperazione nell'ambito del Sistema e allo sviluppo di iniziative di formazione.

La cultura statistica rappresenta uno strumento a disposizione dei soggetti impegnati nella programmazione e nel controllo, disegnando obiettivi misurabili, impegnando a leggere la fenomenologia delle azioni e dei risultati in chiave di nuova progettazione e di nuove strategie.

La sfida della qualità passa non soltanto attraverso la responsabilità degli operatori del Sistema, ma – sempre di più – attraverso il **ruolo dei soggetti** che eseguono le rilevazioni sul territorio, chiamati a sostenere indagini continuative complesse, e a corrispondere a momenti topici come i censimenti. Si tratta di impatti sul territorio di grande rilevanza quantitativa e organizzativa, ma soprattutto di grande portata culturale.

Credo che dovremmo soffermarci sull'esigenza di rendere questi soggetti destinatari di un programma di comunicazione organizzativa che metta al centro dell'attenzione "i valori" più che "le istruzioni". Dagli operatori delle rilevazioni ci si attende affidabilità e autorevolezza, poiché essi sono anche l'interfaccia della statistica ufficiale verso il cittadino; debbono quindi assumere ruoli di chiarificazione, di rassicurazione e promozione. Poiché chiediamo ai cittadini un investimento di tempo, di pazienza e di apprendimento dobbiamo affidare alla rete il compito della spiegazione del "senso" di queste operazioni. Quanto più capillari saranno state le forme con cui avremo costruito e rafforzato l'identità di scopo, tanto più vicino saremo arrivati al cittadino rispondente, acquisendo risposte che saranno date con la convinzione che si sta facendo un investimento nella conoscenza collettiva.

Le opportunità offerte dall'innovazione tecnologica sono un'ulteriore risorsa. Probabilmente ci stanno impegnando più di altre e non è quindi un caso che siano state al centro della nostra riflessione in queste giornate. Ci stanno impegnando e mettendo a dure, seppure stimolanti, prove per la rapidità con cui richiedono la nostra reazione e il nostro adattamento, ma anche per il fatto di presentarsi con una doppia veste: di risorsa per le facilitazioni che prospettano e di

sfida - per il ripensamento al quale ci impegnano in ordine ai paradigmi organizzativi, agli approcci culturali e alle dinamiche operative, di processo e di comunicazione.

4. Le implicazioni delle opportunità tecnologiche

Secondo uno studio recentemente citato da "The Economist", l'investimento in hardware e software per sostenere l'innovazione in tecnologie dell'informazione è pari a un decimo dell'investimento totale. I restanti nove decimi devono essere riversati su tre fattori di sviluppo quali l'innovazione del prodotto, l'innovazione del processo e la formazione del personale.

Il concorso di questi tre fattori genera quello che è stato definito il "capitale organizzativo", che costituisce l'elemento suscettibile di rappresentare il vantaggio competitivo, ovvero lo strumento per il successo di un prodotto o per l'efficacia di un servizio.

Questo quadro di riferimento ci porta a ragionare in una logica di Sistema. Se le opportunità che le tecnologie ci offrono devono essere calibrate rispetto all'organizzazione, non è sostenibile la semplice trasposizione di modelli vincenti in diversi contesti organizzativi.

Poiché siamo Sistema, occorrerà uno sforzo sagace per comprendere quali siano i **valori** da condividere e quali le **innovazioni** da sviluppare, tenendo conto della cultura dei diversi enti.

In ogni caso, per il Sistema statistico nazionale, l'occasione delle tecnologie dell'informazione e la logica della rete offrono una grande possibilità, poiché consentono di rimettere al centro della catena organizzativa il **fattore umano**. Può sembrare una contraddizione, se guardiamo la comunicazione in rete come uno strumento che inibisce il contatto personale; in realtà non lo è, perché la **rete** consente di alimentare situazioni cooperative come ad esempio i **circoli di qualità** nella costruzione del Programma statistico nazionale.

Non è certo una prospettiva semplice, ma se parole d'ordine quali il trasferimento di idee e conoscenze, la solidarietà di Sistema, il rispetto delle culture di ente, l'identità di scopo orientato al servizio finale prevalessero nella consapevolezza degli operatori del Sistema, avremmo l'opportunità di costruire, accanto al capitale organizzativo dei singoli enti, anche un capitale organizzativo di Sistema, che rappresenterebbe il principale motore dello sviluppo.

5. L'applicazione delle tecnologie

Per muoversi nella direzione precedente, l'applicazione delle tecnologie deve essere letta alla luce di due fattori cruciali quali le risorse umane e la comunicazione.

5.1 Le risorse umane

Questa transizione impegna in maniera considerevole il management. E' difficile pensare a un cambiamento organizzativo senza una convinta leadership. La leadership è fortemente esposta; in un sistema a comunicazione spinta e tecnologicamente integrata, gli atti espressi dalla leadership sono immediatamente visibili e valutabili, generano riflessi immediati di comportamento, non soltanto sui collaboratori, ma anche sugli interlocutori esterni all'ente. D'altro canto, in una struttura aperta (che in sé rappresenta un valore primario per le potenzialità comunicative e di interazione), l'efficacia degli atti orientati a dare impulso al cambiamento, risulta potenziata.

I fondamentali classici del management (saper dare fiducia, essere orientati al dialogo, esprimere un impegno sicuro sulla formazione, mirare ai risultati, generare motivazione, insegnare a lavorare insieme e in rete) si pongono perciò al centro del processo di innovazione.

5.2 Gli scenari della comunicazione nel contesto organizzativo

L'applicazione delle tecnologie dell'informazione rende cruciale una strategia della comunicazione all'interno delle organizzazioni.

Ci troviamo di fronte a una dicotomia, rappresentata da una parte dagli straordinari vantaggi offerti dalle tecnologie e dall'altra dai limiti conseguenti dall'uso eventualmente inappropriato degli strumenti. L'uso improprio della posta elettronica, ad esempio, sta rapidamente diventando un fattore di crisi, per la sovrabbondanza di informazioni, per la difficoltà di selezione e soprattutto per la mancanza di protocolli e linee-guida comportamentali nei contesti organizzativi.

Tuttavia, ancora una volta la tecnologia ci viene incontro, ad esempio, offrendo strumenti che si prestano a consentire accessi selezionati.

L'introduzione di software, anche sofisticati, non può prescindere da un lavoro preparatorio della massa documentale. Grandi imprese che hanno avuto l'esigenza di aggiornare rapidamente il personale di altre imprese (aggregate per effetto di fusioni societarie) sui progetti in corso e sulla documentazione relativa, hanno investito in queste operazioni in modo importante, realizzando vere e proprie "stanze elettroniche". L'utenza, anziché ricevere sincronicamente e continuativamente i materiali, ha potuto essere informata degli aggiornamenti e accedere alla documentazione con la certezza di un aggiornamento garantito.

Questo esempio di problematica di comunicazione e di soluzione tecnologica e organizzativa può rappresentare efficacemente la situazione nella quale

ci troviamo, come Istat e come Sistema. Siamo impegnati in uno sforzo di adeguamento organizzativo e dobbiamo investire nella **ricerca di soluzioni** che rappresentino la risposta più adeguata, in un panorama di culture diverse, spinte tecnologiche rapide, domanda dell'utenza e carichi informativi crescenti.

Gli investimenti fatti finora sono stati certamente coerenti.

L'augurio è che, come Istituto e come Sistema, possiamo essere capaci di anticipare strategicamente i prossimi investimenti per assicurare al Paese un'informazione statistica di qualità.