Una valutazione critica dei modelli di accesso remoto nella comunicazione di informazione statistica

M. Lucarelli(*)

(*) ISTAT – Servizio progettazione e supporto metodologico nei processi di produzione statistica

INDICE

1. Introduzione

La crescente domanda di condurre analisi statistiche complesse e particolareggiate su dati individuali, o semplicemente di soddisfare esigenze conoscitive che non trovano riscontro presso i comuni canali di diffusione, richiedendo quindi lo svolgersi di elaborazioni su dati elementari d'indagine, è alla base dell'impegno e della sempre maggiore attenzione che gli Istituti Nazionali di Statistica (INS) ed Eurostat rivolgono a queste tematiche.

Affinché dette esigenze trovino risposta, ed abbia concreta applicazione il principio secondo il quale i dati rilevati costituiscono un patrimonio della collettività, ciascun INS mette a disposizione della comunità scientifica diverse forme di comunicazione dell'informazione statistica. Tra queste, il Data Laboratory rappresenta il servizio che offre maggiore libertà nella conduzione delle ricerche, ma sconta il forte limite di richiedere la frequentazione fisica dei locali ad esso adibiti. Pertanto, grossi sforzi sono stati condotti da diversi INS, al fine di offrire opportunità simili mediante un accesso remoto.

Il lavoro che qui si presenta, propone una riflessione critica sulle soluzioni, sperimentate a livello internazionale, che rappresentano un'evoluzione del servizio offerto dal Data Laboratory, e che in tale contesto si siano affermate come le più diffuse, o le più promettenti.
Nel paragrafo seguente verrà pertanto proposta una classificazione degli strumenti impiegati dai diversi INS nell'ambito della comunicazione dell'informazione statistica. Tale categorizzazione, che non pretende di essere esaustiva, consentirà di definire dei modelli di riferimento che saranno di ausilio nella valutazione comparativa condotta nel paragrafo 3. Nel paragrafo 4 verranno infine discusse alcune considerazioni conclusive.

2. Alcuni modelli di accesso remoto nella comunicazione di informazione statistica

2.1. Elaborazioni personalizzate

Pur non trattandosi di una forma di comunicazione di dati elementari in senso stretto, questo canale di accesso all'informazione statistica, attivo anche presso l'Istat[1], consente di soddisfare quanti abbiano esigenze conoscitive non complesse, ma comunque non previste dai normali piani di diffusione, e che richiedano pertanto l'esecuzione di elaborazioni sui microdati (ad esempio ricercatori che necessitino di particolari tabelle non pubblicate). Gli utenti devono in questo caso illustrare le proprie necessità al personale addetto e contribuire economicamente a sostenere il costo dell'elaborazione. Necessariamente, le elaborazioni non potranno essere troppo complesse, ed il risultato potrà essere disponibile all'utente solo dopo un certo lasso di tempo, dipendente dal fatto che le elaborazioni vengono svolte dallo stesso personale che si occupa dell'indagine.

Il Federal Statistical Office of Germany offre un servizio analogo, ma diversamente strutturato: chiariti gli obiettivi conoscitivi, le analisi vengono svolte dallo stesso personale che si occupa dell'EMAIL-DL (cfr. paragrafo 2.4.1), e pertanto con la possibilità di svolgere analisi anche complesse e in tempi relativamente brevi, sullo stesso patrimonio di dati disponibile presso il Data Laboratory (cfr. paragrafo 2.3).

[1] Per maggiori informazioni, consultare il sito Istat all'indirizzo: http://www.istat.it/servizi/infodati

2.2. File di dati elementari (*Anonymised Microdata Files – AMFs*)

I file di dati elementari (Anonymised Microdata Files, d'ora in poi AMFs), sono collezioni campionarie di dati elementari d'indagine, nei quali l'anonimità delle unità statistiche, ovviamente già prive di identificativi diretti, viene tutelata tramite l'applicazione di diverse metodologie statistiche che, essenzialmente, riducono il contenuto informativo dei dati (*data reduction*), oppure ne alterano il contenuto (*data perturbation*), rendendo in tal modo altamente improbabile la re-identificazione delle unità statistiche (L. Franconi e G. Seri, 2004). Presso l'Istat si adottano soltanto metodi del primo tipo: i file ottenuti, noti come *file standard*, vengono prodotti per diverse indagini dell'Istituto (quasi esclusivamente su individui e famiglie[2]), e sono disponibili a pagamento, previa autorizzazione del Presidente dell'Istat ed adesione ad un contratto che impegna l'acquirente al rispetto della riservatezza dei rispondenti. Altri INS applicano metodi di protezione perturbativi, mentre alcuni applicano metodi che prevedono sia la riduzione che l'alterazione dei dati. Spesso vengono prodotte diverse versioni di AMFs, che hanno contenuti informativi (e quindi livelli di 'protezione') diversi e, conseguentemente, sono destinati a categorie di utenti differenti. In questi casi, sono generalmente disponibili AMFs ad *uso pubblico* (*Public Use Files* – PUFs, anche liberamente scaricabili da Internet, molto protetti e poco informativi), ed AMFs *per la ricerca* (*Licensed Files*, di maggiore contenuto informativo, ed ottenibili previa sottoscrizione di un contratto fortemente vincolante per il ricevente). In alcuni Paesi si costruiscono anche AMFs *ad hoc*, per progetti di ricerca specifici o per raccordare indagini diverse (per maggiori dettagli sulle diverse soluzioni adottate nei vari Paesi si veda: A. Capobianchi, 2005).

2.3. Laboratorio di analisi dei dati (*Data Laboratory – DL*)

Come detto, i file di microdati (AMFs) vengono prodotti riducendo il contenuto informativo dei dati, qualora non vengano impiegate anche metodi *perturbativi*, che ne alterano il contenuto, e/o tecniche di ricampionamento, che diminuiscono il numero di osservazioni. Tali file vengono costruiti cercando di massimizzare la possibilità di impiego da parte dei ricercatori, ma evidentemente non è possibile soddisfare tutte le possibili esigenze di analisi, soprattutto quelle che si dedicano a tipologie di unità statistiche che siano *rare* nella popolazione (e quindi maggiormente a rischio di re-identificazione), o che abbiano necessità di accedere all'insieme completo delle osservazioni rilevate da un'indagine totale.

Le analisi statistiche per le quali è indispensabile avere un dettaglio informativo maggiore di quello disponibile nell'AMF o dati provenienti da indagini totali, possono essere svolte presso i *Data Laboratories* (*DL*). I DL sono uno strumento largamente diffuso presso gli INS europei, e consistono in una o più postazioni di lavoro collocate fisicamente presso una o più sedi dell'Istituto (alcuni Paesi attivano un DL anche presso alcune sedi universitarie). Per accedere al DL, bisogna in genere far parte di un istituto di ricerca o università, sottomettere il proprio progetto di ricerca alla valutazione di un'apposita commissione e, se si riceve l'autorizzazione, sottoscrivere un contratto, che obbliga il ricercatore al mantenimento del segreto statistico. Ottenuta l'autorizzazione, il ricercatore si reca presso il DL e svolge le proprie elaborazioni sui dati elementari messi a disposizione. Terminate le elaborazioni, l'output prodotto viene valutato dagli incaricati del DL (o dalla struttura preposta all'interno dell'INS), e, se ritenuto 'sicuro' sotto il profilo della tutela della riservatezza, rilasciato al ricercatore (in genere via e-mail).
Ad ogni modo, i dati non vengono mai portati al di fuori dei locali del DL, ed ogni forma di output può essere rilasciata solo previa autorizzazione del personale preposto.

Il DL italiano, attivo dal 1999 presso la sede centrale di Roma dell'Istat, prende il nome di "Laboratorio ADELE" (Laboratorio per l'Analisi dei Dati ELEmentari), e mette a disposizione,

[2] La lista completa dei file standard è disponibile presso: http://www.istat.it/servizi/infodati/elenco_file_standard

previa autorizzazione del Presidente, i dati elementari validati delle principali indagini dell'Istituto, senza ridurne né alterarne il contenuto informativo[3].

In altri Paesi, invece, l'INS rende disponibili per l'elaborazione presso i DL solo file protetti (ovvero di ridotto contenuto informativo), o anche perturbati; presso alcuni INS è tuttavia prevista la costituzione di data-set ad hoc, specifici per il singolo progetto di ricerca, ed eventualmente afferenti a diverse indagini. Per lo più i DL sono localizzati sul territorio presso varie sedi dell'INS: solo raramente dispongono di un'unica sede centrale. Inoltre, per autorizzare l'accesso al DL, alcuni INS devono trovare uno specifico interesse nella realizzazione del progetto di ricerca, e taluni richiedono all'utente la sottoscrizione di un contratto che lo equipara al personale interno (vedi Tabella 1).

	Locazione				Dati						
	Capitale	Altre sedi INS	Università	Accesso Remoto	Non alterati	Protetti	Perturbati	Data-set ad hoc	Demo data	Equiparazione dipendenti INS	Necessità interesse INS
Australia	•	•		•		•					
Canada			•	•	•					•	
Danimarca	•			•	•			•	•		
Finlandia	•				•				•		
Germania	•	•		•		•					
Gran Bretagna	•	•		•			•				•
Italia	•				•						
USA	•	•	•	•			•			•	•

Tabella 1: Servizio offerto nei DL di alcuni Paesi[4]

2.4. Accesso remoto (*Remote Access Facilities – RAF*)

Grazie alla diffusione ed alle potenzialità di Internet, è oggigiorno possibile accedere, comodamente dal proprio ufficio o dalla propria abitazione, ad una quantità immensa di informazioni, banche dati e documenti, ed è altresì possibile svolgere le più svariate operazioni, anche qualora queste siano delicate, o coinvolgano dati personali (si pensi, ad esempio, all'*Internet Banking*, ovvero alla gestione del proprio conto corrente bancario via web). Al di là dei problemi tecnici relativi alla sicurezza delle comunicazioni telematiche e dei sistemi informatici coinvolti, che, pur disponendo di varie soluzioni, devono essere affrontati con attenzione, l'utilizzo di Internet nell'accesso all'informazione statistica offre una prospettiva allettante per tutti, e soprattutto per quanti sarebbero attualmente costretti ad effettuare trasferte onerose per condurre le proprie ricerche presso un DL.

[3] Per maggiori informazioni consultare il sito Istat all'indirizzo: http://www.istat.it/servizi/infodati/adele.html
[4] Fonte: A. Capobianchi, 2005 (cit.)

In effetti, la possibilità di offrire al mondo della ricerca scientifica l'opportunità di accedere da postazioni remote alle stesse potenzialità di analisi offerte dai DL, ovvero la formulazione di alternative al DL che non comportino agli utenti l'obbligo di frequentare fisicamente i locali ad esso adibiti, costituisce da tempo un ambizioso obiettivo nel settore della tutela della riservatezza statistica, considerato di prioritario interesse a livello internazionale.

Infatti, diversi INS negli ultimi anni hanno moltiplicato i propri sforzi per raggiungere lo scopo, e molti di essi risultano attualmente dotati di un *RAF* (*Remote Access Facility* – Servizio di Accesso Remoto). Sebbene le soluzioni tecniche ed organizzative adottate nei diversi RAF manifestino caratteristiche peculiari, si possono rintracciare essenzialmente due modelli di riferimento, che nel seguito indicheremo con le sigle *EMAIL-DL* (DL via *e-mail*) e *R-DL* (*Remote* DL).

2.4.1. Il Laboratorio via *e-mail* (*EMAIL-DL*)

L'EMAIL-DL rappresenta la soluzione largamente più diffusa, a livello internazionale, ed è stata adottata dalla quasi totalità dei Paesi che si siano dotati di RAF. Il servizio è essenzialmente basato sull'uso della posta elettronica: tramite e-mail, infatti, gli utenti abilitati mandano al DL un file contenente le istruzioni da far eseguire, e sempre via e-mail, riceveranno l'output. Infatti, la maggior parte dei pacchetti statistici in uso presso i DL offre la possibilità di eseguire in modalità *batch* (cioè senza l'intervento dell'utente) un insieme di istruzioni contenute in un file di testo, e di salvare l'output ed il log delle elaborazioni in appositi file. Sfruttando questa opportunità, è possibile consentire agli utenti di definire le proprie elaborazioni e riceverne l'output senza che essi si spostino dal proprio personal computer. A tal fine, è tuttavia indispensabile fornire agli utenti dei *demo-data*, ovvero data-set che esemplifichino il contenuto e la struttura dei dati sui quali dovranno essere condotte le elaborazioni.

Ovviamente, prima di essere rilasciati all'utente, i risultati delle elaborazioni subiranno una fase di *check* da parte degli addetti, che verificheranno l'assenza di violazioni della privacy.
Questa fase rappresenta l'onere maggiore del servizio per l'INS, soprattutto in termini di risorse umane, ed è scarsamente automatizzabile. In alcune implementazioni[5] si conduce un check automatico 'a priori' sul file di istruzioni, per verificare che non contenga istruzioni il cui output sarebbe certamente non autorizzabile. È tuttavia difficile definire a priori la pericolosità di un'istruzione poiché, in genere, dipende dai dati ai quali sarà applicata, per cui l'insieme delle funzioni disabilitate potrebbe essere eccessivo, oppure troppo ridotto per essere effettivamente utile. In altri casi[6] è stata automatizzata la fase di check sull'output (ma non ne sono specificate le modalità), che comunque, in caso di esito negativo, prevede il controllo manuale. Ad ogni modo, dal momento che l'utente interagisce col DL solamente tramite la posta elettronica, risulta impedita qualsiasi forma di *accesso* ai dati elementari: solo la comunicazione dell'output all'utente resta potenzialmente pericolosa, per cui l'EMAIL-DL si rivela più protettivo del DL ordinario. Va tuttavia considerato che presso questi RAF sono attualmente disponibili quasi esclusivamente dati protetti, quando non perturbati.

[5] Controlli a priori sono condotti presso le Remote Access Facilities delle Agenzie Federali di Statistica statunitensi, ma anche dai servizi sviluppati nell'ambito dei progetti LIS, LES e PiEP.
[6] Ad esempio, presso il Remote Access Data Laboratory (RADL) australiano.

2.4.2. Il *Remote Data Laboratory* (*R-DL*)

Il modello che qui indicheremo con la sigla R-DL, ha maggiori potenzialità e prevede un'architettura informatica più articolata, e finora è stato realizzato solo in Danimarca (L. Borchsenius, 2005) e Svezia (L. J. Söderberg , 2005). In Olanda si sta predisponendo un servizio analogo, mentre in altri Paesi è ancora in fase di studio.
In sostanza, questa soluzione consente all'utente di accedere, da un personal computer predisposto e tramite una connessione Internet, ad applicazioni (in questo caso pacchetti statistici) rese disponibili da un server remoto. Il server si trova fisicamente presso l'INS, e se necessario potrà essere sostituito da un *cluster* di server. Le applicazioni risiedono e vengono eseguite sul server: l'interazione dell'utente con esse viene essenzialmente gestita tramite un comune *browser web*. I dati possono risiedere sul server delle applicazioni, o, meglio, su un server distinto e non accessibile dall'esterno. Agli utenti è consentito l'uso delle applicazioni sui dati messi a loro disposizione, mentre non gli è permesso di salvare alcunché sul proprio computer. In altre parole, gli utenti utilizzano il proprio monitor, mouse e tastiera (quasi) come se si trovassero sulla postazione del DL, mentre in realtà stanno comodamente seduti davanti al proprio computer.
La soluzione adottata da danesi e svedesi prevede l'uso del software commerciale CITRIX[7], che si occupa di '*pubblicare*' le applicazioni (cioè di renderle disponibili da remoto tramite browser web). In Istat soluzioni analoghe sono in fase di sperimentazione per il telelavoro, e, ovviamente, costituiscono una valida prospettiva anche per eventuali evoluzioni del Laboratorio ADELE (cfr. paragrafo 3.5).
L'accesso al servizio è ovviamente preceduto da una fase di autenticazione, che può avvalersi di *token* (come la *RSA SecurId card* in uso presso i R-DL di Danimarca e Svezia) e/o di controlli biometrici (gli olandesi impongono l'uso di un apparecchio che rileva le impronte digitali), ed il traffico di rete viene crittografato in modo da essere al sicuro dalle intercettazioni.

Per quanto riguarda il controllo dell'output da rilasciare agli utenti, le difficoltà restano sostanzialmente quelle del DL: questa architettura permette agli utenti di lavorare dal proprio computer, ma non prevede automazione del processo di *check* delle elaborazioni, che resta quindi manuale ed a carico del personale preposto dall'INS. Nonostante ciò, il modello in uso in Danimarca e Svezia adotta in merito una soluzione originale (per quanto discutibile): gli output vengono spediti automaticamente agli utenti, ogni cinque minuti, tramite posta elettronica. Vengono poi eseguiti dei controlli a campione sugli output spediti e, in caso di violazioni, vengono presi provvedimenti.

Va infine rilevato che l'accesso al servizio offerto in Danimarca e Svezia (ma sembra sarà così anche per quello olandese) può avvenire solo da parte di computer opportunamente configurati da personale dell'INS, per cui l'attivazione di un nuovo progetto di ricerca comporta oneri e problemi tecnici non indifferenti.

[7] http://www.citrix.com

3. Possibili evoluzioni nella comunicazione dell'informazione statistica

3.1. Opportunità e limiti del modello DL: lo spazio per un'evoluzione

In generale, la soluzione del Data Laboratory (DL) come strumento d'accesso ai dati elementari per scopi di ricerca, presenta vantaggi, ma anche limiti.
Gli utenti beneficiano principalmente della libertà di analisi, e del contatto continuo ed immediato col risultato del proprio lavoro: spesso, soprattutto per le analisi più complesse, il ricercatore ha bisogno del *feedback* delle proprie elaborazioni per poterle adattare ai risultati ottenuti (ad es. per la determinazione dei parametri di interesse dei modelli statistici, o anche per la selezione delle variabili da includervi). Per quanto riguarda la nostra esperienza presso il Laboratorio ADELE, si può affermare che solo occasionalmente i ricercatori hanno chiaramente determinate fin dall'inizio le elaborazioni da compiere: piuttosto, queste si delineano tramite approssimazioni successive, esaminando di volta in volta i risultati delle analisi condotte, ed adattando ad essi le elaborazioni. Ad ulteriore conferma di ciò, va considerato il fatto che quasi tutti gli utenti hanno richiesto output parziali, da poter analizzare con calma al di fuori del Laboratorio, ed intervallano spesso le frequentazioni del Laboratorio con pause di riflessione di alcuni giorni, o settimane.
Inoltre, il DL è un ambiente di ricerca massimamente controllato, in cui è possibile rendere fruibili dati che non potrebbero essere altrimenti disponibili (ad es. dati di impresa, di cui solo eccezionalmente vengono prodotti file standard).

Tuttavia, i DL manifestano il loro limite principale nella propria 'fisicità', che costringe tutti gli utenti a condurre le analisi nei locali preposti, con massimo svantaggio per coloro che, provenendo da lontano, devono affrontare trasferte molto costose in termini di tempo e di denaro. Per un ricercatore di Bari o di Milano, ad esempio, recarsi a Roma per una o due settimane per condurre le proprie analisi presenta dei costi non facilmente sostenibili. Infatti, come si evince dal rapporto sull'attività del Laboratorio ADELE (G. Seri e M. Lucarelli, 2004), circa i tre quarti dei responsabili dei progetti di ricerca provengono da Roma, città che ospita l'unico DL italiano, il che manifesta chiaramente l'impatto della collocazione territoriale del DL nei confronti dell'utenza.

Al fine di limitare i disagi indotti da queste trasferte, diversi INS hanno aumentato il numero di DL disponibili sul territorio, localizzandoli presso altre sedi dell'Istituto, o anche presso alcune università. Non va tuttavia trascurato che, in genere, l'impianto di ciascun nuovo DL, a prescindere dal numero di postazioni messe a disposizione degli utenti, comporta per l'INS dei costi non trascurabili, legati al reperimento ed alla messa a norma dei locali, alla fornitura degli elaboratori e del relativo software, ed alla disponibilità di una quantità minima di risorse umane (che dovranno eventualmente essere formate) tale da garantire la continuità del servizio.
Soprattutto, lo sforzo dei diversi INS si è concentrato sull'impiego delle tecnologie informatiche per consentire agli utenti di condurre le proprie elaborazioni da postazioni remote, ovvero senza doversi recare fisicamente presso il DL. Tuttavia, ciascuna soluzione presenta vantaggi ma anche limiti, e la sua applicabilità è fortemente condizionata dalla struttura organizzativa dell'INS e dal contesto giuridico del Paese.

Nel seguito si esamineranno le possibili prospettive di evoluzione del DL ordinario, con un occhio di riguardo alla loro applicabilità nella realtà italiana e tenendo in considerazione l'esperienza finora condotta tramite il Laboratorio ADELE.

3.2. Soluzione EMAIL-DL

Il modello concettuale che si è qui indicato come EMAIL-DL, prevede sostanzialmente che gli utenti comunichino via e-mail i propri programmi (i.e. insieme di istruzioni scritte nel linguaggio di un determinato pacchetto statistico) al personale dell'INS, e che ricevano via e-mail il risultato delle proprie elaborazioni.
Tale modello ha come pregio la semplicità di realizzazione, tale da sembrare immediata: al di là di complicazioni tecniche di entità del tutto trascurabile, relative alla riservatezza delle e-mail, la messa in opera del servizio in forma minimale prevede la sola attivazione di una casella di posta elettronica.
Per di più, dal momento che gli utenti non avrebbero in alcun modo accesso ai dati, il modello appare anche più protettivo per la tutela della riservatezza di quanto non lo sia lo stesso DL, in cui i ricercatori hanno comunque la possibilità di vedere a video i dati elementari.

Il fascino di una soluzione concettualmente semplice che possa rispondere all'esigenza degli utenti di condurre le proprie elaborazioni senza recarsi fisicamente presso il DL, può indurre a sottovalutarne gli aspetti negativi , che invero non sembrano trascurabili.

Anzitutto, come argomentato nel paragrafo precedente, presso il DL il ricercatore si avvale della possibilità di avere una interazione dinamica ed immediata con le proprie elaborazioni, vedendone direttamente a video i risultati. Tale interazione, che costituisce un pregio del servizio offerto dal DL, essendo impedita nel modello dell'EMAIL-DL, ne rappresenta un rilevante limite: in questo senso, gli utenti avrebbero un servizio peggiore rispetto al DL classico.
Al di là dell'eventuale insoddisfazione degli utenti, che potrebbe peraltro essere bilanciata dall'opportunità di lavorare dal proprio computer, il limite ora evidenziato si tradurrebbe soprattutto in un aumento notevole del carico di lavoro per il personale incaricato del rilascio dell'output. Infatti, è lecito supporre che gli utenti avrebbero bisogno, per poter proseguire il lavoro, dell'output di *ciascuna* esecuzione dei loro programmi (ovvero di *tutte* le elaborazioni intermedie, anziché del solo output finale), e ciò andrebbe ad appesantire inutilmente la parte del processo che, richiedendo l'intervento umano, rappresenta il costo maggiore per l'Istituto. Tale valutazione si aggrava se consideriamo che l'eliminazione del vincolo di doversi recare presso il DL comporterebbe, con buona probabilità, una moltiplicazione del numero delle richieste.

Un altro aspetto del modello dell'EMAIL-DL che grava sul personale dell'Istituto è la necessità di mettere a disposizione degli utenti i cosiddetti *demo-data*. Tali data-set sono indispensabili per consentire agli utenti di predisporre i programmi, devono essere identici nella struttura ai dati che vengono messi a disposizione per l'elaborazione, e rispecchiarne il più possibile le caratteristiche statistiche, senza ovviamente replicarne il contenuto. Ciò implica che, *per ciascuna indagine* che si intende mettere a disposizione dell'utenza presso l'EMAIL-DL, è necessario condurre un'approfondita analisi sulle caratteristiche statistiche dei dati, decidere quali di queste sia opportuno mantenere nel data-set simulato, ed infine predisporre il data-set simulato. Quest'attività potrebbe essere svolta dal personale afferente all'indagine in questione, oppure potrebbe andarsi ad aggiungere a quelle in carico al personale del DL, ma un coinvolgimento dei responsabili d'indagine sarebbe comunque indispensabile.

Infine, se da un lato è vero che gli utenti possono evitare di recarsi presso il DL, è altrettanto necessario che vi sia del personale preposto che mandi in esecuzione i programmi, anche se in parte l'impatto di questo aspetto può essere limitato, cercando di imporre agli utenti di usare nei propri programmi standard sintattici che consentano l'automazione almeno parziale di questa parte del processo.

In sostanza, l'applicazione del modello dell'EMAIL-DL richiede un investimento in termini di risorse umane tale da scoraggiarne l'applicazione, senza considerare che tali risorse dovrebbero peraltro essere altamente specializzate nell'ambito della tutela della riservatezza statistica.

3.3. Soluzione Elaborazioni Personalizzate

Secondo il modello utilizzato in Germania, il servizio delle Elaborazioni Personalizzate potrebbe essere considerato come una forma di sviluppo del DL ordinario. In effetti, tale servizio è già attivo presso il nostro Istituto, ma è strutturato in modo diverso, ed offre opportunità di analisi molto limitate rispetto a quelle possibili in Germania. Secondo tale modello, sarebbe prevista una forte interazione col ricercatore committente, il quale dovrebbe precisare esattamente l'output che intende ottenere, mentre la fase dell'elaborazione sarebbe esclusivamente a carico del personale del DL. La fase di controllo della riservatezza dell'output sarebbe in questo caso ridotta al minimo, in virtù della approfondita fase istruttoria del progetto di ricerca.

Tuttavia, valgono anche in questo caso le considerazioni fatte nel paragrafo 3.1, riguardo al fatto che assai raramente un progetto di ricerca ha ben delineato fin dall'inizio il risultato che intende ottenere e, quindi, le elaborazioni necessarie. In questo senso, sarebbe un servizio che potrebbe accontentare solo quanti richiedano statistiche piuttosto semplici, come tabelle o indicatori, ma non chi abbia bisogno di complessi modelli statistici. Inoltre, la necessità di far comprendere a terzi le proprie esigenze, e dover attendere che questi compiano le elaborazioni, comporta per gli utenti una sicura dilatazione dei tempi di esecuzione del progetto di ricerca, anche se in parte bilanciata dal non essere soggetti alla frequentazione fisica del DL.

Ovviamente, sarebbe anche necessario un serio investimento in termini di risorse umane, le quali dovrebbero affiancare l'utente nella lunga fase di definizione preliminare del progetto, e dovrebbero poi svolgere concretamente le elaborazioni. Per di più, sarebbe richiesta loro una competenza specifica sulle singole indagini prodotte dall'Istituto che, data la numerosità e la varietà delle indagini condotte, risulterebbe di difficile reperimento.

In sostanza, questa soluzione potrebbe essere vista come un servizio aggiuntivo al DL ordinario, ma sembra che i costi in termini di risorse umane eccedano i benefici per l'utenza, che sarebbero comunque a favore di pochi.

3.4. Soluzione R-DL

Nonostante non rappresenti la soluzione più diffusa, il modello del DL remoto offre potenzialità tali da renderlo una delle prospettive più interessanti e moderne per l'evoluzione del servizio offerto dal Laboratorio ADELE.
Questa soluzione si basa sull'impiego di strumenti di *Remote Desktop*, ovvero software in grado di trasferire il desktop di un sistema (o di generarne uno apposito) su un altro elaboratore. Con le opportune cautele in ambito di sicurezza informatica, è possibile in questo modo consentire ad un utente di lavorare, dal proprio computer e tramite internet, proprio come se si trovasse su una postazione del DL, senza tuttavia imporgli di recarsi fisicamente presso i locali del DL.
Il modello dell'R-DL risulta quindi particolarmente interessante, dal momento che risolve il problema principale del DL ordinario (la collocazione fisica delle postazioni) ma non limita l'interazione dei ricercatori con l'elaboratore come avviene con l'EMAIL-DL.

Dal momento che gli utenti lavorano proprio come se fossero presso il DL, dal punto di vista dell'organizzazione del lavoro, il modello in esame non porta modifiche significative, a parte, ovviamente, l'impianto del sistema di connessione ed autenticazione tramite opportuni strumenti informatici. Pertanto, come nel DL ordinario, la fase istruttoria per l'autorizzazione dei progetti di ricerca e la fase di controllo dell'output finale delle elaborazioni richiede l'intervento del personale addetto.
In Danimarca, gli utenti del servizio ricevono automaticamente, ogni cinque minuti, i risultati delle proprie elaborazioni: i controlli vengono effettuati a campione, ed a posteriori. L'automazione della

fase di rilascio, in questo caso, raggiunge lo scopo di ridurre il carico di lavoro sul personale addetto, ma viene effettuata a scapito del controllo effettivo dell'output, e si appoggia di fatto alle norme contrattuali che vincolano l'operato degli utenti. In pratica, anziché impedire che gli utenti possano acquisire output non autorizzato, ci si fida del loro operato, salvo controlli sporadici. Tale fiducia è presumibilmente basata sulla convinzione che le norme di accesso al servizio prevedano sanzioni o svantaggi tali da scoraggiare gli utenti a compiere violazioni. A prescindere dalla validità di questa considerazione, bisogna comunque tener presente che gli utenti non sono, in genere, esperti di tutela della riservatezza statistica, e possono compiere violazioni pur essendo in buona fede. Anzi, per quanto riguarda l'esperienza condotta nel Laboratorio ADELE, risulta in genere difficile far comprendere agli utenti quale output non corrisponda ai vincoli di tutela della riservatezza, e perché.

Queste osservazioni introducono l'aspetto maggiormente delicato del modello R-DL, ovvero la questione dell'autenticazione dell'utente, ed il grado di fiducia che deve essergli attribuito. Nel DL ordinario, l'utente, per accedere alla postazione, si presenta di persona agli addetti del DL, e se ne può quindi accertare l'identità. Inoltre, il suo operato è (almeno potenzialmente) costantemente sotto controllo da parte del personale del DL. Dal punto di vista informatico, si possono agevolmente definire sistemi di autenticazione ragionevolmente sicuri, ma sono per lo più basati su credenziali che è difficile impedire che vengano (volontariamente) trasferite tra le persone. Ad esempio, si potrebbe dare al ricercatore autorizzato una *userid* ed una *password*, da usare assieme ad un *token* (ad es. un dispositivo hardware da connettere al computer), ma nessuno impedisce al ricercatore di cedere queste credenziali a terzi. E' per questo che in Olanda viene imposto l'uso di un dispositivo che rileva le impronte digitali. D'altra parte, ancora una volta, la difficoltà si può superare tramite l'imputazione di responsabilità formali nei confronti dell'utente: se il sistema di credenziali è robusto, si presume che chiunque le usi abbia l'autorizzazione dell'utente legittimo, e che quindi agisca in sua vece; pertanto, eventuali illeciti sarebbero comunque sotto la responsabilità dell'utente al quale era stato originariamente conferito l'accesso al servizio.
Per quanto riguarda il controllo sull'operato degli utenti, la situazione è ancora più delicata, e fa comunque affidamento sul fatto che l'utente rispetti i vincoli contrattuali che regolano il servizio. Infatti, gli utenti vedono sul monitor del proprio computer i dati elementari, oltre agli output parziali eventualmente a rischio, e non c'è modo di impedire, ad esempio, che prendano nota di alcuni valori, o che fotografino lo schermo.

3.5. Soluzione R-DL in ambiente controllato: il DL Regionale

Le difficoltà ascrivibili all'autenticazione degli utenti e al controllo del loro operato, nell'ambito del modello R-DL, sembrano difficilmente risolvibili, se non imponendo la sottoscrizione di norme sanzionatorie scoraggianti, e fidandosi che ciò sia sufficiente.
Tali problemi non si porrebbero, però, se il computer da cui si accetta la connessione si trovasse in un ambiente controllato da personale di fiducia.
Per inciso, va detto che il servizio offerto in Danimarca, Svezia e Olanda non prevede mai la connessione da un computer qualsiasi (né personale), ma solo da computer selezionati all'interno dell'azienda o ente presso cui l'utente presta servizio; il personale dell'INS deve predisporre tali computer, e deve mantenere libero accesso ad essi. Pertanto, al servizio offerto da questi Paesi non si può accedere da qualsiasi computer, ed impone tempi di accesso non brevi, oltre a causare trasferte di personale dell'INS.

In tale contesto appare plausibile una soluzione ibrida: laddove l'INS disponga di sedi dislocate sul territorio nazionale, si potrebbe consentire l'accesso al servizio dell'R-DL solo a postazioni in esse collocate, risolvendo così le difficoltà del modello R-DL sopra evidenziate.
Tale soluzione, che indicheremo come DL Regionale, è particolarmente idonea alla realtà italiana, che può disporre di uffici regionali, e consente di moltiplicare sul territorio i punti di accesso del DL

ordinario mantenendo i costi entro limiti accettabili. Finora, infatti, attivare una nuova sede del DL equivaleva a replicare in loco l'intera struttura del DL, richiedendo quindi un notevole investimento, soprattutto in termini di risorse umane. Il modello del DL Regionale, invece, prevede un'architettura distribuita, ma fortemente centralizzata: la struttura effettivamente impiegata dall'erogazione del servizio resterebbe sempre la sede centrale del DL, mentre presso le sedi locali si dovrebbe solamente consentire l'accesso degli utenti alla postazione dedicata.

Il servizio manterrebbe quindi le caratteristiche positive del DL (libertà di analisi per i ricercatori, massima sicurezza per la tutela dei dati personali) configurandosi quindi come un miglioramento del servizio offerto dal DL ordinario, senza tuttavia imporre costi di realizzazione e gestione che possano essere gravosi per l'Istituto.

3.6. Soluzione *Virtual Data Laboratory* (V-DL)

Da ultima, consideriamo l'ipotesi del Laboratorio Virtuale (*Virtual Data Laboratory* – V-DL*)* come una potenziale forma di sviluppo del DL ordinario (una sperimentazione in merito è documentata in: S. Polettini *et al.*, 2006). Dal momento che si tratta dell'ipotesi di un servizio non ancora implementato da alcun Paese, il modello del V-DL non è stato presentato nel paragrafo 2, ma verrà brevemente descritto nel seguito.

L'idea alla base del V-DL consiste nel mettere a disposizione degli utenti un'interfaccia web tramite la quale siano disponibili, ad utenti autenticati, un insieme limitato di funzioni e di dati, tali per cui l'output delle elaborazioni sia certamente al riparo da possibili violazioni della riservatezza. La valutazione dell'output sarebbe in questo caso effettuata automaticamente, attingendo ad un insieme di regole aprioristiche, che possono dipendere dal tipo di elaborazione richiesta, dai dati sui quali viene svolta l'elaborazione, ed anche dall'output stesso. A puro titolo d'esempio, si potrebbe decidere di autorizzare tabelle a doppia entrata contenenti percentuali su un'indagine sociale e non su altre indagini, ma solo se le percentuali (ovvero le celle della tabella) derivano da almeno tre unità statistiche. Queste regole, ma anche i dati e le funzioni, potrebbero essere archiviate su un database, al fine di essere aggiornabili e facilmente disponibili al sistema, che quindi potrebbe accrescersi continuamente di contenuti.
L'architettura informatica di una soluzione del genere sarebbe costituita: (i) da un server web, che mette a disposizione l'interfaccia; (ii) da una logica di controllo implementata in un linguaggio di scripting (PHP[8], ad esempio, che è *open source*); (iii) da un motore statistico (si potrebbe usare R[9], anch'esso open source) che compie le elaborazioni, oltre (iv) al già citato database che potrebbe contenere le regole per il rilascio, i dati ed eventualmente anche i metadati disponibili per l'indagine. All'utente non sarebbe richiesto di imparare il linguaggio del pacchetto statistico, dal momento che la sua interazione sarebbe solo con l'interfaccia web: sarà compito della logica di controllo associare alle scelte effettuate dall'utente (tramite click del mouse) un insieme di istruzioni e parametri opportunamente assemblati da passare al motore statistico. Fondamentalmente, le regole definirebbero un insieme di percorsi possibili che l'utente potrebbe effettuare tramite l'interfaccia, stabiliti in modo tale che l'output delle elaborazioni, che rappresenta il punto finale di ciascun percorso, sia con certezza rilasciabile. Pertanto, essendo la fase di controllo effettuata automaticamente e a priori, i risultati delle elaborazioni potrebbero essere visualizzati immediatamente sul browser degli utenti, ed anche resi disponibili per il download in locale.

Il V-DL apparterrebbe pertanto alla categoria di sistemi nota come "*enabling systems*", ovvero sistemi in cui sono abilitate solo alcune operazioni; in contrapposizione, la categoria dei "*disabling systems*" prevede che alcune funzioni vengano disabilitate, e ciò è quanto avviene negli EMAIL-DL

[8] http://www.php.net
[9] http://www.r-project.org

che prevedono un filtro sui programmi mandati dagli utenti. In merito, va notato come il compito di disabilitare tutte le funzioni potenzialmente pericolose di un pacchetto statistico ricco di comandi e di opzioni sia certamente più gravoso e di risultato meno sicuro rispetto all'individuazione di alcune funzioni statistiche da ritenere sicure, senza considerare che in genere gli EMAIL-DL mettono a disposizione più di un pacchetto statistico, mentre una soluzione come il V-DL potrebbe adottarne uno solo.

La soluzione del V-DL ha il grosso vantaggio di essere totalmente automatizzata, giacché evita il ricorso al controllo manuale dell'output da parte del personale. Inoltre, essendo tutti gli output producibili al riparo da violazioni della riservatezza statistica, anche l'accesso al sistema potrebbe essere conferito via Internet e senza troppi vincoli (al limite potrebbe essere liberamente disponibile a tutti, anche se ciò sembra eccessivo), mentre attualmente l'accesso agli AMFs o ai DL è limitato ai soli fini di studio e di ricerca, ed è subordinato alla sottoscrizione di contratti vincolanti per l'utente.
D'altra parte, tale automazione ha un costo iniziale alto: al di là dell'attivazione dell'infrastruttura informatica, che comunque prevede anche l'implementazione della logica di controllo e non può quindi essere immediata, la difficoltà maggiore è rappresentata dall'individuazione delle regole decisionali, che implicano uno sforzo di generalizzazione statistico-metodologico sulla 'pericolosità' degli output delle singole statistiche, in quanto devono condurre ad asserzioni di validità generale non ancora documentate in letteratura.

A riguardo, tuttavia, devono essere considerati due aspetti.

Anzitutto, la modularità del sistema consente di aggiungere funzioni nel tempo: ciò vuol dire che non è necessario che tutte le statistiche che si intende rendere disponibili a regime debbano essere presenti fin dall'inizio, né che debba essere delineato con precisione l'insieme finale delle statistiche da offrire agli utenti. In sostanza, si potrebbe attivare il sistema anche con una sola statistica, ed aggiungerne altre nel tempo. Ovviamente, ciò vale anche per i dati.

In secondo luogo, non tutte le statistiche sono ugualmente interessanti per gli utenti: evidentemente, un sistema come quello in discussione dovrebbe prioritariamente offrire statistiche e modelli di vasta applicabilità e di interesse generale; modelli di impiego specifico avrebbero un costo di implementazione troppo alto rispetto ai potenziali utilizzatori. Ne consegue che il V-DL, non potendo soddisfare le necessità di analisi più complesse o particolari, potrebbe affiancare, ma comunque non sostituire, il Data Laboratory tradizionale.

Per quanto riguarda le richieste giunte in questi anni al Laboratorio ADELE, circa la metà riguardavano la produzione di tabelle, un quarto intendevano condurre regressioni, e circa il 15 % erano interessate ad indicatori.
La tematica della tutela della riservatezza nell'ambito della produzione e diffusione di tabelle statistiche è da anni oggetto di studio, e pone problematiche tali da non consentirne l'impiego in un sistema generalizzato quale quello del V-DL. Infatti, per stabilire se una data tabella è rilasciabile, bisogna tener conto anche delle altre tabelle già rilasciate e derivanti dallo stesso insieme di dati: ciò sottintende un sistema che abbia memoria storica degli output rilasciati e possa considerarli nella valutazione dell'output corrente (il che, oltre a causare una rapida saturazione dell'insieme delle tabelle rilasciabili, contrasta con il modello del V-DL, che prevede la valutazione 'a priori' dell'output, basata su regole generalizzate), oppure implica un sistema in cui sia preventivamente e staticamente definito l'insieme delle tabelle rilasciabili (ovvero, sostanzialmente, il piano di diffusione dell'indagine, la cui definizione evidentemente esula dalle prerogative del V-DL).
Viceversa, sistemi automatizzati che consentano di stimare modelli di regressione (noti come *model server*) sono già stati realizzati (V. Gambhir e K.W. Harris, 2005), e rappresentano esperienze positive ed incoraggianti (P. Steel e A. Reznek, 2005). Anche la costruzione di indicatori, altra categoria di output largamente richiesto dagli utenti, potrebbe essere consentita tramite il V-DL, purché di struttura non troppo complessa e calcolati su un numero sufficientemente ampio di unità statistiche.

In conclusione, il modello del V-DL non sembra certamente adatto a soddisfare tutti gli utenti del DL tradizionale, ma potrebbe essere di grande ausilio per quanti abbiano bisogno di risultati che, per le caratteristiche stesse della statistica di interesse, siano sicuri sotto il profilo della tutela della riservatezza. Inoltre, un servizio del genere, affiancato al DL ordinario, consentirebbe di annullare, almeno per una parte dei progetti di ricerca, la fase burocratica preliminare relativa all'autorizzazione e la fase di verifica dell'output, ovvero di automatizzare una parte del lavoro che attualmente necessita dell'intervento umano, mentre gli utenti avrebbero il vantaggio di evitare attese per l'autorizzazione o il rilascio dei dati e, soprattutto, potrebbero condurre l'elaborazione via web, dal proprio computer.

4. Conclusioni

Tra le soluzioni di evoluzione delle forme di comunicazione dell'informazione statistica qui presentate, quella del R-DL appare particolarmente promettente, in virtù delle potenzialità del servizio offerto agli utenti e della relativa semplicità di realizzazione. Essa tuttavia pone dei problemi formali e sostanziali, nell'ambito dell'autenticazione e del controllo dell'operato degli utenti.

Il modello del DL Regionale, che rappresenta un'applicazione del R-DL presso ambienti controllati, risulta di notevole interesse, soprattutto per l'alta applicabilità alla realtà italiana, in quanto fornisce un miglioramento netto del servizio offerto dal DL ottenibile in tempi brevi e con costi contenuti.
Anche il modello del V-DL, nell'accezione di *model server*, pone una prospettiva di evoluzione molto interessante e che sarebbe senz'altro gradita all'utenza, ma necessita di impegnativi approfondimenti metodologici in merito, e pertanto non si configura come una soluzione a breve periodo.

Sembrano invece meno attraenti le ipotesi del EMAIL-DL e di evoluzione del servizio delle Elaborazioni Personalizzate, perché da un lato offrono un servizio più limitato rispetto al DL ordinario, e dall'altro impongono all'Istituto notevoli investimenti, soprattutto in termini di risorse umane qualificate.

BIBLIOGRAFIA

L. Borchsenius, "New developments in the danish system for access to micro data", invited paper presentato presso la "Joint UNECE/Eurostat work session on statistical data confidentiality", Geneva, Switzerland, 9-11 November 2005, disponibile on line presso: http://www.unece.org/stats/documents/ece/ces/ge.46/2005/wp.2.e.pdf

A. Capobianchi, "Alcune esperienze in ambito internazionale per l'accesso ai dati elementari", Documenti Istat n.8/2005

L. Franconi e G. Seri (a cura di), "Metodologie e tecniche di tutela della riservatezza nel rilascio di informazione statistica", Istat, collana Metodi e Norme, n. 20 – 2004

V. Gambhir e K.W. Harris, "ANalytical Data Research by Email and Web (ANDREW)", supporting paper presentato presso la "Joint UNECE/Eurostat work session on statistical data confidentiality", Geneva, Switzerland, 9-11 November 2005, disponibile on line presso: http://www.unece.org/stats/documents/ece/ces/ge.46/2005/wp.7.e.pdf

S. Polettini, A. Ponti, M. Lucarelli, M. D'Alò, F. Solari "Stimatori per piccole aree su web: un'esperienza open source", in B. Liseo, G.E. Montanari, N. Torelli (a cura di) "Metodi statistici per l'integrazione di dati da fonti diverse", FrancoAngeli, 2006

G. Seri e M. Lucarelli: "Il Laboratorio per l'analisi dei dati elementari (ADELE): monitoraggio dell'attività dal 1999 al 2004", Documenti Istat n.9/2004

L. J. Söderberg, "MONA – Microdata ON-line Access at Statistics Sweden", invited paper presentato da presso la "Joint UNECE/Eurostat work session on statistical data confidentiality", Geneva, Switzerland, 9-11 November 2005, disponibile on line presso: http://www.unece.org/stats/documents/ece/ces/ge.46/2005/wp.3.s.e.pdf

P. Steel e A. Reznek, "Issues in designing a confidentiality preserving model server", invited paper presentato presso la "Joint UNECE/Eurostat work session on statistical data confidentiality", Geneva, Switzerland, 9-11 November 2005, disponibile on line presso: http://www.unece.org/stats/documents/ece/ces/ge.46/2005/wp.4.e.pdf