Documenti Istat

Alcune esperienze in ambito internazionale per l'accesso ai dati elementari

A. Capobianchi (*)

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

Sommario

La crescente richiesta da parte di ricercatori di utilizzare dati individuali per analisi sempre più specifiche, ha creato, per i vari Istituti di statistica, l'esigenza di sviluppare metodologie di accesso e tecniche di protezione che, da una parte soddisfino al meglio tali richieste e dall'altra garantiscano il rispetto del principio della tutela della riservatezza dei rispondenti. A tale scopo si sono sviluppati in particolare tre canali dedicati:
Rilascio di dati elementari, Accesso remoto e Laboratori di analisi di dati.
I singoli Istituti di statistica hanno sviluppato, e di conseguenza caratterizzato, i tre canali in base alle normative legate alla tutela del diritto alla riservatezza vigenti nel proprio paese.
Nel presente documento si descrivono le soluzioni adottate da alcuni paesi, che rappresentano in maniera significativa l'attuale stato di avanzamento delle tecniche e delle metodologie che rendono accessibili dati individuali.

Indice

1. Introduzione

Le analisi statistiche su dati individuali stanno diventando sempre di più uno strumento necessario nel dare assistenza alla messa a punto di programmi e nella valutazione di decisioni. Le indagini statistiche forniscono sempre maggiori informazioni per la conoscenza della società e la disponibilità di dati collezionati ed analizzati da agenzie predisposte possono fornire nuove opportunità di studio e di analisi del comportamento sociale ed economico della società stessa. Da ciò deriva una maggiore richiesta sia di accesso a collezioni di dati sia di una qualità sempre migliore dei dati rilasciati. Sebbene gli Istituti nazionali di statistica raccolgano una grande mole di dati, la loro diffusione deve confrontarsi costantemente con il principio della tutela della riservatezza.
I problemi inerenti il rilascio e l'accesso a dati elementari vengono affrontati dai diversi Istituti di statistica sia nell'ambito della ricerca di nuove tecniche di protezione dei dati che nell'ambito della definizione di nuovi canali per l'accesso ai dati stessi.
Qui di seguito daremo una breve classificazione dei canali maggiormente utilizzati come mezzo per l'accesso ai dati elementari dai vari Istituti di statistica.

Rilascio di dati elementari : I dati elementari vengono rilasciati sotto forma di file di dati individuali trattati attraverso l'applicazione di una o più tecniche di protezione in modo tale che non sia possibile, tramite l'utilizzo di mezzi ragionevoli, l'identificazione delle unità rilevate.
Tali file vengono generalmente rilasciati su supporto informatico per un utilizzo esterno all'Istituto di statistica. E' possibile definire diversi livelli di protezione dei file in base ai vincoli posti al loro utilizzo e alla loro circolazione . In generale distinguiamo tra:

- *file di dati accessibili al pubblico*; il livello di protezione applicato rende estremamente improbabile l'individuazione dei rispondenti; il rilascio non comporta la sottoscrizione di un contratto con l'Istituto che li produce; in alcuni casi sono disponibili su siti Web.
- *file di dati accessibili per motivi di ricerca*; il livello di protezione applicato rende il rischio di identificazione di unità rilevate molto basso (Per una trattazione sistematica di tale argomento si rimanda a Istat, 2004) ; il rilascio avviene previa sottoscrizione di un contratto che regola il comportamento che l'utente deve avere nel trattare i dati contenuti nel file; il contenuto informativo in file di questo tipo è maggiore rispetto a quello dei file di microdati accessibili al pubblico.

Accesso Remoto: L'idea di base che caratterizza questo canale di diffusione è quella di permettere ad utenti esterni di eseguire in remoto, per mezzo di una rete informatica, delle analisi su file di microdati senza avere un accesso diretto ad essi. Tale accesso può essere fornito in due modi:

- *Esecuzione remota*: l'utente invia un file contenente codici di programmazione per eseguire delle analisi dei dati. Tali programmi vengono eseguiti più o meno automaticamente da personale dell'istituto. La restituzione dell'output al richiedente avviene solo dopo aver verificato che non ci siano violazioni della riservatezza. L'operazione di verifica può essere fatta sfruttando alcuni pacchetti statistici, accettando ad esempio solo particolari comandi (*filtering software for set-up*) o manualmente da personale dell'Istituto. La restituzione dell'output avviene dopo un tempo che varia da qualche minuto a qualche giorno.
- *I Laboratori Virtuali* sono dei siti web in cui gli utenti hanno la possibilità di effettuare analisi statistiche su microdati in tempo reale.

Attraverso l'accesso remoto è possibile analizzare dei dati più dettagliati rispetto a quelli contenuti nei file rilasciati tramite CD-Rom o disponibili via Web. Ciò è realizzabile in quanto i dati restano all'interno dell'Istituto nazionale di statistica ed è possibile effettuare dei controlli sulle elaborazioni eseguite. Un'ulteriore vantaggio di questo canale è legato al fatto che risulta essere un mezzo molto duttile per l'accesso ai dati. In effetti sono diversi i modi in cui viene implementato a seconda sia

del paese che dei dati a cui è applicato. Tale differenze risulteranno evidenti in sede di analisi delle esperienze nei paragrafi successivi.

Laboratori di analisi dei dati : Sono dei locali protetti, collocati all'interno degli Istituti di statistica o altre istituzioni, dove gli utenti, previa autorizzazione, possono analizzare dati elementari. In tali locali personale dell'Istituto di statistica controlla, dal punto di vista della riservatezza, sia l'intero processo di analisi che l'output prodotto.

Per rendere più chiara l'esposizione dei successivi paragrafi diamo qui di seguito delle brevi definizioni delle tecniche di protezioni maggiormente applicate durante il processo di controllo del rischio di identificazione dai vari Istituti di statistica; (per una trattazione sistematica di tale argomento si rimanda a Istat, 2004). In generale le tecniche di protezione vengono classificate in tecniche non perturbative e tecniche perturbative. Tra le prime ricordiamo:

Ricodifica globale: consiste nell' unione di classi adiacenti per variabili qualitative e nella riduzione in classi per variabili continue. Casi particolari di ricodifica sono i cosidetti *top-coding* e *bottom-coding* che coinvolgono rispettivamente le code a destra e a sinistra della distribuzione della variabile da proteggere.

Soppressione locale: consiste, per particolari unità statistiche rilevate, nella sostituzione del valore osservato con il codice relativo al valore mancante in una o più variabili.

Eliminazione di unità: consiste nell' eliminazione di alcune unità selezionate.

Eliminazione di variabili: non vengono rilasciate le informazioni relative ad alcune variabili che possono essere considerate o altamente identificative o troppo riservate.

Tra le tecniche perturbative abbiamo:

Data swapping: la protezione si ottiene scambiando in un insieme di coppie di record un sottoinsieme delle variabili (Istat, 2004).

Perturbazione casuale dei valori o Aggiunta di disturbo: questa tecnica consiste nell'aggiunta di un disturbo casuale ai valori originali (Istat, 2004).

La differenza sostanziale tra le due tipologie di tecniche sta nel fatto che, mentre per le tecniche perturbative il controllo del rischio si ottiene attraverso una vera e propria modificazione dei dati, per quelle non perturbative il controllo si ottiene attraverso una riduzione del contenuto informativo preservando quindi l'integrità dei dati stessi

Nelle sezioni successive verranno analizzate le esperienze di alcuni paesi che rappresentano in maniera significativa l'attuale stato di avanzamento delle tecniche e dei metodi per rendere accessibile l'informazione statistica.

2. Canada (*Statistics Canada*)

Per poter permettere un maggiore utilizzo delle proprie collezioni campionarie l'Istituto nazionale di statistica canadese ha sviluppato, esclusivamente per motivi di ricerca, diversi canali di accesso ai dati elementari che garantiscono la riservatezza dei rispondenti alle indagini (Tambay *et al*., 2003).
I canali messi a disposizione dall'Istituto sono:

- file di dati elementari (Public Use Microdata File – PUMS);
- accesso remoto;
- laboratori di analisi dei dati.

2.1. File di dati elementari (PUMFs)

Statistics Canada rilascia file di microdati protetti (Public Use Microdata File - PUMFs) per motivi di ricerca già dai primi anni '70[1]. Il maggiore sviluppo nella diffusione di tali file si è avuta nel 1996 con l'introduzione del "Data Liberation Iniziative" (DLI)[2] (ovvero un accordo, tra le principali università e Statistics Canada, che regola l'accesso e la diffusione ai principali prodotto statistici tra i quali i file PUMFs.
La maggior parte dei PUMFs rilasciati i file sono relativi ad individui o a famiglie (file individuali o gerarchici) e ad ognuno di essi viene allegata una documentazione metodologica relativa al disegno campionario e alle variabili contenute nel file stesso. Sono invece raramente rilasciati i file di dati di impresa o longitudinali in quanto, per tali dati, si ritiene che sia troppo elevato.
L'autorizzazione al rilascio di microdati, da parte di una commissione apposita, si ha solo se tale rilascio comporta un effettiva valorizzazione dei dati dell'indagine e se vengono applicate tutte le procedure necessarie affinché non sia possibile l'identificazione dei rispondenti.
Per quanto riguarda i metodi di protezione Statistics Canada applica una vasta gamma di tecniche perturbative e non[3]:

- perturbazione casuale dei valori,
- data swapping,
- sostituzione di valori critici con valori medi,
- tecniche di imputazione,
- soglie sulla dimensione di sottopopolazioni critiche,
- ricodifica globale,
- eliminazione di record o variabili dal file.

Le informazioni geografiche sono date solo a livello aggregato e molte delle informazioni circa il disegno campionario (strati e cluster) non vengono rilasciate. Nel caso in cui per motivi di riservatezza non viene rilasciata la variabile relativa ai pesi campionari, per molti PUMFs vengono forniti dei coefficienti di variazione che permettono di calcolare delle "grossolane" stime delle varianze per semplici statistiche come la media, proporzioni, rapporti o differenze tra rapporti. Nel caso di variabili categoriche i coefficienti di variazione vengono calcolati considerando l'effetto del disegno sulla formula della varianza per campionamenti casuali semplici senza reinserimento e considerando un insieme di valori conservativi come il 75° percentile ecc. (Tambay *et al*., 2003)

[1] http://www.statcan.ca/english/Dli/continuumofaccess.htm, 6-2005 data dell'ultima visita al sito.
[2] http://www.statcan.ca/english/Dli/whatisdli.htm, 6-2005.
[3] http://www.statcan.ca/english/freepub/12-539-XIE/steps/disclosure.htm, 6-2005.

2.2. Accesso Remoto

Già nei primi anni 90' Statistics Canada propone l'accesso remoto[4] come canale di accesso indiretto ai dati proprio per sopperire a quelle richieste di informazioni non soddisfatte dai PUMFs: ad esempio particolari variabili non contenute nei PUMFs stessi o indicazioni campionarie necessarie per il calcolo esatto della varianza.
In un primo momento l'accesso remoto è stato offerto solo per un piccolo numero di indagini campionarie, per alcune delle quali non era prevista la predisposizione di file PUMFs. L'utilizzo di tale canale è ristretto ad un numero molto limitato di ricercatori, i quali devono presentare un dettagliato progetto di ricerca. Solo dopo l'accettazione di tale progetto e previa sottoscrizione di un contratto che regoli l'uso dei dati e degli output viene rilasciata l'autorizzazione all'accesso ai dati.
Nel caso in cui è predisposto un accesso di tipo remoto, devono essere fornite tutta una serie di informazioni necessarie all'utente per poter formulare i proprio programmi di analisi. In particolare: una documentazione campionaria che includa la descrizione della struttura del file di dati confidenziali (file di base o file master); il file di prova che segua la struttura del file master affinché l'utente possa testare i propri programmi su tale file; un meccanismo che permetta di ricevere programmi via e-mail e successivamente di inviare indietro l'output dell'elaborazione dopo un accurato controllo manuale dal punto di vista della riservatezza. Un controllo di tipo automatico risulterebbe troppo difficile sia a causa della possibilità, da parte degli utenti, di utilizzare più pacchetti statistici quanto per la complessità dei dati trattati.
Di seguito descriviamo gli approcci seguiti per l'accesso in remoto da due indagini differenti (Tambay *et al*., 2003): National Population Health Survey (NPHS) e Survey of Labour and Income Dynamics (SLID)[5].

1) National Population Health survey (NPHS)

L'indagine longitudinale NPHS è l'indagine che maggiormente è stata richiesta attraverso il canale dell'accesso remoto. Il successo di tale indagine è sicuramente legato alla completezza delle informazioni disponibili. Infatti, è a disposizione degli utenti una documentazione dettagliata sul campionamento, un dizionario sui dati, un file sintetico molto realistico su cui gli utenti possono testare i propri programmi e la possibilità di utilizzare diversi software come SPSS, SAS, Stata ecc. ed infine risultano relativamente brevi i tempi di ritorno degli output.
Il file sintetico viene accuratamente predisposto per replicare la struttura del campione di base anche se contiene un numero inferiore di record.

2) Survey of Labour and Income Dynamics (SLID)

Anche questa indagine è di tipo longitudinale ed è caratterizzata da una struttura molto complessa. Tale complessità è legata alle particolari unità di analisi incluse nell'indagine. Per agevolare l'accesso e l'utilizzo di tali dati si è sviluppato un particolare sistema di reperimento dati (SLIDRET). Attraverso questo sistema l'utente può creare un dataset che corrisponde alle proprie necessità senza dover comprendere l'intera struttura dei dati.
Un utente che voglia procedere all'analisi di tali dati attraverso il canale dell'accesso remoto, deve per prima cosa creare un proprio file di analisi attraverso il sistema SLIDRET e, successivamente costruire i file di programma (SAS, SPSS, STATA) da mandare via e-mail a Statistics Canada che per mezzo di personale autorizzato provvederà ad eseguirli sul file di analisi. Esiste una versione pubblica di SLIDRET alla quale è associato una database con una struttura vuota.

[4] http://www.statcan.ca/english/edu/rda/index.htm, 6-2005.
[5] http://www.statcan.ca/english/research/75F0002MIE/75F0002MIE1994014.pdf, 6-2005.

I file di output ottenuti dopo aver eseguito le analisi richieste dall'utente vengono infine analizzati per garantire la riservatezza e se soddisfano questo requisito vengono restituiti via e-mail all'utente.

2.3. Laboratorio di analisi dei dati

Il programma di sviluppo del Laboratoro di analisi dei dati è parte di una iniziativa che coinvolge l'Istituto nazionale di statistica, il Consiglio di ricerca sulle Scienze Sociali e Umane (SSHRC), alcune Università Canadesi e la Fondazione Canadese per l'Innovazione con lo scopo di aiutare e fornire supporto alla ricerca sociale canadese[6] .
I laboratori di analisi dei dati forniscono l'accesso a dati confidenziali all'interno di ambienti protetti. Tali Laboratori si trovano in nove università Canadesi, selezionate dal Consiglio di ricerca sulle Scienze Sociali e Umane e accuratamente controllate dal punto di vista della sicurezza da parte di Statistics Canada. Per accedere ai laboratori è necessario presentare un dettagliato progetto di ricerca che viene accuratamente analizzato da una commissione apposita costituita da personale dell'Istituto di statistica e del SSCHR. I ricercatori, il cui progetto viene approvato, devono sottoscrivere un contratto che, dal punto di vista della riservatezza dei dati, li rende equiparabili ai dipendenti di Statistics Canada. Vengono quindi messi a disposizione esclusivamente i dati specificati nel progetto che possono essere elaborati utilizzando diversi pacchetti statistici disponibili all'interno dei laboratori come GAUSS, ISML, SPSS,SAS, STATA e SHAZAM. Infine, affinché venga garantita la riservatezza dei rispondenti, gli output prodotti che si vogliono rimuovere dal laboratorio, vengono accuratamente analizzati da dipendenti di Statistics Canada che comunque svolgono il loro lavoro di controllo durante tutto il processo di analisi.

[6] http://www.statcan.ca/english/rdc/index.htm, 6-2005.

3. Danimarca (*Statistics Denmark*)

Prima di descrivere come l'istituto di statistica danese abbia sviluppato specifiche procedure di accesso a file di microdati è bene ricordare che, come per altri paesi nord europei, la maggior parte della produzione statistica è basata su registri dello popolazione e delle società[7]. In particolare Statistics Denmark organizza i suoi dati in cosiddetti registri statistici che sono costituiti al solo scopo di definire l'insieme delle statistiche correnti[8]. In generale tali registri si basano su *registri amministrativi* che vengono costruiti e mantenuti da più uffici pubblici e che quindi collezionano dati per scopi differenti da quelli puramente statistici. Ad esempio il registro statistico della popolazione si basa per la maggior parte sul registro centrale della popolazione istituito nel 1968 e gestito dal ministero degli interni danese.
Nella costruzione dei registri un particolare ruolo è svolto dalla *chiave identificativa*; la regola fondamentale è che ad ogni singola unità registrata (persona fisica o impresa economica) viene associato un identificativo che poi viene utilizzato come chiave. L'unicità della *chiave identificativa* può essere considerata la vera forza di questo particolare sistema di rilevazione; in effetti un corretto uso delle *chiavi identificative* permette di effettuare collegamenti tra diversi registri permettendo così di ottenere una enorme offerta di informazione su cui fare ricerca. Ad esempio per un individuo (o una generica unità di rilevazione) è possibile ottenere delle informazioni nel tempo (analisi longitudinale) o, collegando informazioni contenute in più registri, è possibile monitorare una vasta gamma di informazioni che coprono più caratteristiche dello stesso individuo (o unità).
A metà degli anni '80 l'Istituto nazionale di statistica Danese ha cominciato a verificare, da parte di enti di ricerca e divisioni ministeriali, un crescente interesse a collezioni di dati individuali da analizzare per scopi di ricerca. Tale crescente richiesta di dati però non poteva essere soddisfatta dall'Istituto a causa delle regole sulla riservatezza dei dati stabilite dalla dirigenza dello stesso istituto e a causa delle leggi esistenti sulla divulgazione dei dati da registri.
Nell'ottica di una migliore e maggiore regolamentazione della diffusione di microdati nel luglio del 2000 viene approvata la legge sull'*Elaborazione dei Dati Personali* che riguarda l'intero processo di trattamento e diffusione di tali dati. Questa legge implementa inoltre una direttiva dell'Unione europea (Direttiva 95/46/EC) sulla diffusione ed elaborazione di dati personali all'interno dell'Unione Europea.
L'Istituto nazionale di statistica ha così stabilito un cosiddetto "*sistema di adattamento per ricercatori esterni"* offrendo la possibilità di analizzare dati individuali tramite i seguenti canali (Andersen, 2003):

- Laboratorio di dati elementari;
- Sistema di accesso in remoto;
- File di microdati per motivi di studio.

3.1. Laboratorio di analisi dei dati

Nel 1986 è stato istituito il primo laboratorio di dati elementari all'interno dei locali dell'Istituto nazionale di statistica Danese e in questo laboratorio, ricercatori autorizzati possono aver accesso a dati individuali estratti dai vari registri statistici. Essenzialmente l'Istituto di statistica mette a disposizione degli utenti (o ricercatori) un data-set creato sulla base delle informazioni contenute nella descrizione del progetto di ricerca che si vuole condurre. Il laboratorio viene amministrato centralmente dalla Divisione di Ricerca e Metodi, e proprio lo staff di tale divisione si occupa della creazione di gran parte dei data-set interdisciplinari richiesti. Tale attività richiede in alcuni casi

[7] http://www.dst.dk/HomeUK/ForSale/Research/acces.aspx , 6-2005.

[8] http://www.dst.dk/HomeUK/ForSale/Research/demographic.aspx?, 6-2005.

molto lavoro in quanto spesso i progetti di ricerca proposti richiedono un insieme di dati la cui creazione coinvolge diversi registri statistici o problemi di definizioni non sempre banali. Per ridurre i costi che devono essere affrontati per la creazione di data-base ad *hoc* e per risolvere particolari problemi su alcuni dati specifici l'Istituto di statistica Danese ha predisposto la creazione di alcuni data-base, detti data-base di ricerca, che coprono i settori specialistici maggiormente richiesti dai diversi utenti .
Il data-base più frequentemente utilizzato è quello integrato per le ricerche sul Mercato del Lavoro (IDA) istituito nel 1991 sulla base della collaborazione dell'Istituto nazionale di statistica e il Consiglio delle Ricerche sulle Scienze Sociali Danesi. Altri data-base di ricerca di notevole rilievo sono il Data-base Demografico, il Data-base sulla Fertilità ecc.. Alcuni di essi sono stati costruiti direttamente su richiesta di alcuni istituti di ricerca che hanno anche contribuito economicamente alla loro costruzione.
I dati contenuti nei database messi a disposizione dei ricercatori nel Laboratorio hanno lo stesso dettaglio informativo dei dati contenuti nei registri a parte per i codici identificativi che vengono logicamente rimossi e sostituiti con una chiave fittizia. Tali dati possono essere elaborati liberamente anche se qualsiasi processo che permetta di identificare individui o imprese deve essere escluso. I dati elementari rimangono all'Istituto di Statistica Danese mentre i risultati statistici aggregati, che comunque garantiscono il rispetto delle regole sulla riservatezza, vengono rilasciati all'utente il quale può utilizzarli per analisi successive.
Il ricercatore e tutti i partecipanti al progetto di ricerca firmano un contratto con l'Istituto con il quale garantiscono il rispetto delle regole sulla riservatezza e tutti i manoscritti e le pubblicazioni devono essere sottomesse preventivamente all'Istituto di Statistica per un controllo che verifichi il rispetto delle regole stesse.
A partire dal 1996 sono stati fatti importanti investimenti per migliorare la qualità del servizio reso dal laboratorio di analisi attraverso l'acquisizione di una piattaforma Unix, miglioramenti sulla capacità dei computer stessi ed inoltre mettendo a disposizione degli utenti una più vasta gamma di programmi di analisi statistica. Sono così disponibili, oltre al programma SAS, programmi come SPSS, STATA e GAUSS.
L'Istituto nazionale di statistica Danese, per permettere ai vari utenti di sviluppare sulle proprie postazioni programmi da applicare successivamente ai dati all'interno del laboratorio, ha creato dei piccoli data-set che coinvolgono al massimo 1000 record e che possono essere prestati, dietro richiesta, ai ricercatori. In generale comunque i data-set creati a questo scopo sono molto pochi e coprono solo una piccola parte dei campi di ricerca possibili.

3.2. Accesso remoto

Nel 2000 l'Istituto di Statistica Danese ha istituito una commissione con il compito di valutare la possibilità di utilizzare un sistema di accesso ai microdati che prevedesse l'elaborazione degli stessi direttamente dalle postazioni degli utenti ovvero di un sistema del tipo *accesso remoto*.
E' stato così proposto uno schema per cui gli utenti di nazionalità danese, previa speciale autorizzazione, possono accedere in remoto a pre-determinate collezioni di dati sulle quali vengono condotte le elaborazioni previste dal progetto di ricerca in esame. In un primo momento solo alcune collezioni campionarie erano rese disponibili tramite l'accesso remoto; in particolare venivano esclusi tutti i dati relativi alle imprese ed altri ritenuti particolarmente sensibili come quelli giudiziari. Inoltre, se la richiesta riguardava l'intera popolazione, le variabili venivano preventivamente limitate.
Dal 2002 lo schema di acceso in remoto è stato equiparato all'accesso ai dati tramite laboratorio in termini di garanzia della riservatezza. Come conseguenza è possibile accedere tramite accesso remoto a tutti i dati disponibili nel laboratorio, ovvero è possibile consultare dati relativi all'intera popolazione e con lo stesso dettaglio presente nei registri statistici mantenuti dall'Istituto.

Il rendere disponibile una tale quantità e qualità di dati anche tramite accesso remoto ha reso necessario la definizione di nuove regole per l'approvazione delle autorizzazioni all'accesso dei dati stessi. In particolare l'autorizzazione all'accesso ai microdati può essere data esclusivamente per scopi di ricerca quindi a ricercatori appartenenti ad università, istituti di ricerca, istituti ministeriali o organizzazioni di ricerca per scopi umanitari. Per ottenere tale autorizzazione deve essere presentato un progetto di ricerca che viene attentamente valutato dagli organi preposti dell'Istituto di Statistica Danese. Nel caso in cui il progetto viene accettato si sottoscrive un contratto con l'Istituto con il quale si assicura il rispetto di alcune regole predisposte per garantire la riservatezza dei dati. Come per il laboratorio, il data-base di analisi è predisposto dal personale della Divisione di Ricerca e Metodi e viene collocato su un particolare server Unix utilizzabile esclusivamente dagli utenti esterni. L'accesso in remoto al server è permesso solo alle persone autorizzate dall'Istituto alle quali viene assegnata una speciale password. Tutto il processo di analisi dei dati viene quindi svolto sul server che si trova all'interno dell'Istituto Danese. Gli utenti possono elaborare i dati creando anche dei nuovi data-set ma non è possibile effettuare nessun trasferimento dei dati dal server alla propria postazione. I risultati delle analisi vengono salvati in una particolare area del server e successivamente inviati via e-mail agli utenti stessi. L'invio dei risultati avviene in tempi relativamente brevi (circa ogni 5 minuti) e le e-mail contenenti gli output inviati vengono salvate in una area apposita. Solo successivamente alcuni output vengono scelti in maniera casuale e controllati dal personale predisposto che, se verifica una qualche forma di violazione della riservatezza, prontamente contatta l'utente in questione.

3.3. File di dati elementari

L'Istituto non rilascia in generale file di dati elementari. Unica eccezione sono i "data-set di studio". Tali insiemi di dati sono stati creati, a partire dal database integrato sul Mercato del Lavoro. Lo scopo è quello di offrire agli studenti la possibilità di analizzare ed applicare su dati reali modelli statistici di analisi del mercato del lavoro e in campo sociologico. Questi data-set coinvolgono poche migliaia di individui che vengono seguiti nel tempo (data-base longitudinale) su un ristretto insieme di variabili. I dati contenuti in questi file non vengono modificati quindi le caratteristiche fondamentali dei dati sono preservate.

4. Finlandia (*Statistics Finland*)

Come per la maggior parte dei paesi nord-europei, anche in Finlandia la maggioranza dei micro-dati statistici deriva dai registri amministrativi. Anche se tali dati possono essere generalmente diffusi per scopi di ricerca la Legge Statistica Finlandese (1994) stabilisce che qualsiasi dato ottenuto per scopo statistico è riservato indipendentemente dalla fonte da cui deriva. Quindi vengono considerati riservati sia i dati relativi ad indagini campionarie condotte dall'Istituto Finlandese sia quei dati derivanti da fonti amministrative nel momento in cui sono in possesso dell'Istituto.
Per permettere di analizzare dati individuali si sono sviluppati i seguenti canali[9]:

- Laboratori di dati elementari;
- Sistema di accesso in remoto.

4.1. File di dati elementari

In Finlandia il rischio di identificazione per informazioni personali viene controllato attraverso delle restrizioni sui dati piuttosto che, come nel caso della Danimarca, imponendo delle limitazione sui luoghi in cui vengono messi a disposizione (laboratori analisi).
Vengono quindi rilasciati file di dati elementari che non includono informazioni particolarmente sensibili e tali per cui sia possibile identificare i singoli individui solo attraverso un notevole sforzo sia di mezzi che economico. In particolare i dati diffusi sono dati campionari i cui codici identificativi ufficiali vengono sostituiti con dei codici fittizi, vengono eliminate tutte le informazioni identificative dirette (nome cognome indirizzo ecc..), alcune variabili considerate particolarmente riservate vengono completamente rimosse ed alcune ricodificate a livelli più aggregati. Tale tipo di rilascio riguarda esclusivamente dati relativi ad individui e famiglie.

4.2. Laboratorio di Analisi dei dati

Un diverso trattamento è invece riservato ai dati di impresa. Essendo tali dati molto più difficili da proteggere attraverso le usuali tecniche di protezione applicate ai dati personali, Statistics Finland ha predisposto l'accesso ad essi solo attraverso il laboratorio di analisi dei dati istituito all'interno dei locali della sede dell'Istituto stesso. Tale laboratorio, che si trova nella capitale, è stato istituito nel 2001 con l'obiettivo principale di rendere accessibili i dati di impresa. All'interno del Laboratorio i dati sono organizzati in diversi file sas che sono tra loro collegabili per mezzo di pseudo identificatori. In questo modo i ricercatori possono costruirsi in maniera indipendente gli insiemi di dati che rispondono alle esigenze del progetto di ricerca che si vuole svolgere.
Per agevolare i ricercatori, in particolare quelli che vivono lontano dalla capitale, vengono predisposti dei file test o demo-data sui quali è possibile condurre delle analisi preliminari e quindi elaborare programmi che successivamente vengono applicati ai dati reali che si trovano all'interno del laboratorio.
Gli output prodotti vengono infine controllati, dal punto di vista della riservatezza, da personale dell'Istituto.

4.3. Accesso remoto

L'Istituto di Statistica Finlandese non prevede la possibilità di analizzare dati elementari in accesso remoto.

---

[9] http://www.isi-2003.de/guest/3558.pdf?MItabObj=pcoabstract&MIcolObj=uploadpaper&MInamObj=id&MIvalObj=3558&MItypeObj=application/pdf 6-2005.
http://www.micro2122.scb.se/Access_to_microdata_in_the_Nordic_countries.pdf 6-2005.

5. Germania (*Federal Statistical Office of Germany*)

Prima degli anni '70 i dati aggregati che normalmente venivano prodotti e pubblicati dal German Federal Statistics Office venivano considerati un prodotto sufficiente ed esaustivo per soddisfare le richieste fatte dai diversi utilizzatori di dati ufficiali. Nei primi anni settanta un gruppo di ricercatori universitari richiedono, per la prima volta, al German Federal Statistics Office una collezione di microdati da utilizzare in un progetto di ricerca che aveva l'obiettivo di creare un sistema di indicatori sociali e politici. Proprio questo studio mette in evidenza un cambiamento nelle richieste degli utenti che sempre più frequentemente richiedono file di microdati per studi econometrici e per applicazioni a modelli multivariati possibili grazie ad una evoluzione sempre più veloce di sistemi hardware e software.
Nel 1981 viene approvata la prima legge (legge Statistica Federale)[10] che regola la diffusione di dati ufficiali sotto forma di microdati. Tale legge prevede la diffusione di dati in forma individuale *totalmente anonimizzata*, definiti come dati da cui non è possibile ottenere in alcun modo informazioni a livello di singola unità rilevata. Il livello di protezione applicato ai cosiddetti "PUF" (*Public Use file*) è molto alto e ciò comporta una notevole perdita di informazione. Proprio lo scarso contenuto informativo di questi file non favorisce la diffusione dell'utilizzo degli stessi in particolar modo nel mondo scientifico. Nel 1987 viene riconosciuto il cosiddetto "Privilegio della Scienza" e la nuova versione della legge Statistica Federale prevede la possibilità dell'utilizzo, da parte di ricercatori dei vari istituti di ricerca, di file di dati anonimizzati *di fatto*. Tali file prendono il nome di "*Scientific Use file*" (SUF) e il rischio di identificazione ad essi associato non è nullo ma il costo, in termini di mezzi tempo e denaro, da sostenere per ottenere un'identificazione è sproporzionato. Logicamente il contenuto informativo dei SUF è nettamente maggiore rispetto a quello dei PUF. Ulteriori miglioramenti in termini di possibilità di accesso e qualità dell'informazione diffusa si sono avuti con l'istituzione del primo *Research Data Centre* (RDC) in Wiesbaden nell'ottobre del 2001. Nel 2002 si è poi stabilita la creazione di un centro per ogni stato federale. I *Research Data Centre* offrono diverse opportunità di accesso ai microdati:

- creazione di PUF e SUF
- creazione di file per studenti a scopo di studio (*Campus File*)
- accesso in sede di microdati attraverso le cosiddette *Safe Scientific Workstation*
- accesso remoto (*Controlled Remote Data Processing*

5.1. File di dati elementari - Public Use File (PUF) Scientific Use File (SUF) Campus File

Una prima possibilità di utilizzare microdati è quella di acquistare file PUF o SUF su cd-rom. Abbiamo già notato come la differenza tra i due tipi di file è data dal livello di protezione applicata e quindi dal diverso contenuto informativo dei file stessi. Per ottenere un file protetto vengono utilizzate le tecniche di ricodifica globale, sottocampionamento ed eliminazione di dati critici. Nel caso particolare di dati economici, per i quali la protezione risulta più difficoltosa, viene applicata anche la tecnica di soppressione locale che purtroppo comporta una notevole perdita di informazione.
I PUF sono file che possono essere acquistati ed utilizzati da tutti mentre i file SUF sono rivolti solo a ricercatori tedeschi. Quindi ricercatori stranieri non possono utilizzare dati SUF ma possono comunque accedere ai microdati attraverso altri canali.

I Centri di Ricerca predispongono inoltre dei speciali file, detti "*Campus File*" con lo scopo di sviluppare la cultura statistica a livello Universitario e scolastico. Tali file sono completamente

[10] http://www.unece.org/stats/publications/statistical.confidentiality.pdf

anonimizzati e possono essere scaricati liberamente da internet. La differenza sostanziale con i PUF, anch'essi completamente anonimizzati, è legata al contenuto informativo. Infatti i Puf sono creati per scopo di ricerca e quindi anche se protetti rispecchiano le informazioni statistiche contenute nei file da cui vengono prodotti. I Campus file invece hanno il solo scopo di poter utilizzare dati nell'insegnamento statistico quindi non è detto che rispecchino la realtà empirica. Per questo sono prodotti in maniera semplice e molto veloce.

5.2. Accesso Remoto - Controlled Remote Data Processing (CRDP) Special Data Processing (SDP)

Nel caso in cui un ricercatore necessiti di maggiori informazioni di quelle contenute nei file PUF e Suf o dati relativi a indagini campionarie per le quali non sono stati creati dei file protetti i Centri di Ricerca Dati hanno sviluppato due ulteriori canali, il Controlled Remote Data Processing (CRDP) e lo Special Data Processing (SDP).
Nel Caso del CRDP si può procedere come segue. In un primo momento il ricercatore può condurre le sue analisi sui dati anonimizzati o su file di dati detti "strutturali". Questi ultimi sono dei file che risultano essere uguali a quelli contenenti i dati originali in termini di struttura ma non in termini di contenuti. In questo modo vengono prodotti programmi di analisi in formato SAS, SPSS, STATA, ecc.. che successivamente vengono inviati ai Centri di Ricerca Dati. Qui personale apposito provvede all'applicazione degli stessi ai dati i quali sono protetti in misura inferiore rispetto ai SUF o, in alcuni casi, non protetti.
Una forma speciale di accesso remoto è il cosiddetto "Processo Speciale sui Dati". In questo caso il ricercatore descrive il progetto di ricerca ad un rappresentante del Centro che successivamente svolge, in maniera autonoma, l'intero processo di analisi dei dati.

5.3. Laboratorio di Analisi dei Dati – Safe Scientific Workstaion (SSW)

Anche in Germania è possibile accedere ai microdati attraverso dei Laboratori di Analisi situati all'interno degli stessi Uffici di Statistica. I ricercatori possono accedere ai dati attraverso workstation dedicate predisposte all'interno di detti laboratori.
Su tali computer vengono messi a disposizione file di dati comunque protetti anche se ad un livello inferiore rispetto a quelli distribuiti come SUF. Gli output (in generale sottoforma di tabelle) vengono rilasciati previo controllo da parte di personale dell'Istituto appositamente addetto.
In passato era possibile accedere ai dati non protetti attraverso uno speciale canale detto *"One-Dollar-man"* . Ovvero il ricercatore firmava un contratto con l'Istituto attraverso il quale diventava a tutti gli effetti un dipendente dell'Istituto stesso. La condizione di "dipendente" permetteva al ricercatore di analizzare i microdati non protetti rispettando le regole di riservatezza come un qualsiasi altro dipendente dell'Istituto. Tale procedura non viene più applicata in quanto viene considerata una soluzione eccessiva.

6. Gran Bretagna (*Official National Statistics – ONS*)

L'Istituto nazionale di statistica britannico è responsabile della produzione della statistica ufficiale del paese. In Gran Bretagna non esiste una legge statistica unica che regoli il rilascio e l'accesso ai microdati. Una regolamentazione in tal senso può trovarsi all'interno della legge che disciplina la raccolta dei dati o è definita in base alla particolare tipologia dei dati che si vogliono analizzare. La legge più importante relativa al tipo di informazione che può essere analizzate è la legge sulla protezione dei dati (the Data Protection Act 1998) che contiene le regole generali sull'analisi di dati individuali. Nel 2002 la dirigenza dell'ONS ha approvato la creazione di una procedura centralizzata per l'autorizzazione all'accesso e al rilascio di dati individuali, il Protocollo per l'Accesso ai Dati e la Riservatezza. La Procedura per il Rilascio dei Microdati (Microdata Release Procedure – MRP) è stata successivamente approvata nel gennaio del 2003.
Tutte le richieste per l'accesso ai dati vengono valutate da un commissione di esperti. L'ONS approva la richiesta solo se è certa che alcuni principi fondamentali definiti dalla Protocollo di accesso ai dati sono soddisfatti[11].
In generale ai microdati vengono applicate le usuali tecniche di protezione[12] tra le quai ricordiamo anche la tecnica di perturbazione casuale dei valori [13]. Per quanto riguarda i canali di diffusione messi a disposizione dall'Istituto sono:

- File di dati individuali;
- Accesso in remoto;
- Laboratorio di dati elementari.

6.1. File di dati elementari

Affinché un file dati elementari *(Anonymised microdata file AMFs)* venga rilasciato deve essere preventivamente sottoposto alla commissione di esperti che verifica il rispetto delle regole sulla riservatezza definite nel Protocollo per l'Accesso ai dati. E' possibile rilasciare dei file di microdati anche come *file pubblici (AMFPs)*. Ciò avviene quando i dettagli territoriali, il disegno campionario e il limitato contenuto informativo del file è tale da garantire la riservatezza dei rispondenti. Tali file non vengono prodotti per scopi di ricerca ma esclusivamente per permettere delle analisi limitate o per esercitazioni universitarie.
Alcuni file di dati elementari possono essere rilasciati esclusivamente dietro contratto (*AMF for Licensed used AMFLs*). Questo avviene quando la riservatezza del file può essere garantita solo ed esclusivamente se essi vengono utilizzati per particolari scopi di ricerca e quindi da utenti selezionati.
Di particolare importanza sono i file di dati elementari relativi al censimento del 2001 (*Sample of Anonymised Records SARS*); esistono tre tipi di file SARS.
Il file individuale al *tre per cento* contiene più di 1,5 milioni di record, le informazioni individuali sugli argomenti principali relativi al censimento ed inoltre alcune informazioni riassuntive circa la famiglia di appartenenza delle singole unità presenti nel file.
Il file familiare gerarchico all' *uno per cento* permette di collegare tutti gli individui appartenenti alla stessa famiglia, le informazioni territoriali sono rilasciate a livello di contea e contiene informazioni relative a circa 216.000 famiglie.

---

[11] http://www.statistics.gov.uk/about/NS_ONS/ONS_microdata_releases.asp, 6-2005.

[12] http://www.statistics.gov.uk/about/data/methodology/general_methodology/sdc.asp, 6-2005.

[13] http://www.statistics.gov.uk/events/rss_ons_conf/downloads/, 6-2005.

Infine viene predisposto il file SARs al *cinque per cento per piccole aree* che prevede un dettaglio territoriale molto elevato (Tranmer, *et al.,* 2005). Tale dettaglio territoriale va a discapito di altre variabili che vengono rilasciate ad un livello più aggregato rispetto agli altri due tipi di file.
Tutti e tre tipi di file SARs sono stati protetti attraverso tecniche di perturbazione dei dati che vengono applicate alle variabili considerate più a rischio o più identificative.

6.2. Accesso remoto (Remote Settings)

La Procedura per il Rilascio di Microdati (MRP) prevede l'accesso a microdati confidenziali anche attraverso il canale dell'accesso remoto. L'utilizzo di tale canale deve essere preventivamente approvato dalla commissione per il rilascio dei microdati la quale deve attentamente valutare il progetto di ricerca prima di autorizzare l'accesso ai dati stessi. Una volta ottenuta l'autorizzazione è possibile inviare i propri programmi di analisi via internet.

6.3. Laboratori di analisi dei dati (Safe Centres)

I "Safe Centres" sono laboratori di accesso a dati individuali dove, previa autorizzazione della Commissione per il Rilascio dei Microdati, è possibile analizzare file di microdati confidenziali.

Per quanto riguarda l'accesso ai dati di impresa il "*Business data Linking Project*"[14] (BDL) fornisce l'accesso a microdati di impresa attraverso un laboratorio sicuro, localizzato negli uffici di Londra dell'ONS, dove ricercatori universitari possono svolgere le proprie analisi statistiche su dati di impresa che altrimenti non sarebbero consultabili. L'accesso è riservato esclusivamente a ricercatori di Università o Enti di ricerca e per il momento non esistono facilitazioni per studenti o dottorandi. L'accesso al Laboratorio è vincolato all'approvazione del progetto di ricerca, i cui risultati devono coincidere con gli interessi dell'ONS, e alla sottoscrizione di un contratto tra il ricercatore e l'ONS.

Delle versioni più dettagliate dei file SARS individuale all'uno per cento e SARS familiare al cinque per cento sono accessibili in laboratori di analisi (Conrolled Access Microdata Sample - CAMS) collocati nelle sedi dell'ONS di Londra, Titchfield, Newport, Southport e prossimamente in Edinburgh e Belfast[15]. Per ottenere l'accesso a tali dati deve essere sottoposto il progetto di ricerca all'approvazione di una apposita commissione (Census Research Access Board – CRAB). Una volta approvato il progetto viene sottoscritto un contratto che regola le condizioni di accesso e il comportamento da seguire nell'utilizzo dei dati. All'interno dei laboratori sono messi a disposizioni diversi programmi come SPSS, STAT, Excel ecc. L'output prodotto viene controllato da personale dell'ONS per garantire che soddisfi i principi di riservatezza definite nelle linee guida delle politiche di rilascio dei dati. Così come i file SARs disponibile su CD-Rom anche i file messi a disposizione nei laboratori sono protetti tramite tecniche perturbative.

[14] http://www.statistics.gov.uk/about/bdl/default.asp, 6-2005.
[15] http://www.statistics.gov.uk/census2001/downloads/confidentiality_guidelines.doc, 6-2005.
http://www.statistics.gov.uk/census2001/sar_cams.asp, 6-2005.

7. Australia (*Australian Bureau of Statistics – ABS*)

Il rilascio e l'accesso di file di microdati è diventato, per l'Australian Bureau of Statistics (ABS) uno degli elementi più importanti nella definizione della strategia di diffusione di informazione statistica. L'importanza del rilascio di dati nella forma di record individuali è legata da una parte alla sempre maggiore richiesta da parte del mondo scientifico di tale forma di output da utilizzare per l'analisi empirica di modelli ecc.., dall'altra dalla necessità dell'Istituto stesso di incoraggiare e seguire progetti di ricerca in ambito sociale.
Se per l'Istituto il rilascio di microdati sempre più accurati è da una parte uno degli obiettivi da raggiungere, dall'altra il garantire la riservatezza degli intervistati è fondamentale per mantenere la fiducia dei rispondenti ai censimenti e alle varie indagini campionarie svolte dall'Istituto stesso.
Nel 1981 una Commissione parlamentare redige un documento in cui si promuove e si auspica un miglioramento nella legislazione statistica che promuovi la massima diffusione e utilizzazione dei dati disponibili. Nel 1983 questa avvertenza prende forma nella tredicesima sezione della legge 1905 sui Censimenti e la Statistica (Census and Statistics act 1905) che da la possibilità all'Istituto di statistica di diffondere informazioni in forma di record statistici individuali *non identificabili*. I file di microdati che, attraverso l'applicazione di diverse tecniche di protezione, vengono resi non identificabili prendono il nome di file di record unitari confidenzializzati (Confidentialised Unit Record Files - CURFs[16]). Il rilascio, in qualsiasi forma, di file CURF è altamente controllato; una commissione apposita (Microdata Review Panel) è stata istituita per analizzare le varie proposte di diffusione dei dati. In particolare tale commissione verifica se il file CURF predisposto garantisce i rispondenti da possibili identificazioni "spontanee" o identificazioni causati da possibili collegamenti tra il file in esame e dati pubblici o in possesso del richiedente i dati.
In generale le tecniche applicate dall'Australian Bureau of Statistics per proteggere file di microdati sono[17]:

- ricodifica globale
- perturbazione casuale di valori
- data swapping
- eliminazione di alcuni record dal file

mentre per quanto riguarda i canali di diffusione messi a disposizione dall'Istituto sono:

- CD-Rom
- Accesso in remoto
- Laboratorio di dati elementari.

Da ricordare che, qualsiasi sia il canale di diffusione prescelto, i dati messi a disposizione di utenti esterni sono sempre dati anonimizzati ovvero file CURF. Vengono prodotti tre tipi di file CURF: *base*, *esteso* e *speciale* che differiscono tra loro dal grado di protezione applicato.

7.1. File di dati elementari[18]

Il livello di protezione applicato ai file CURF rilasciati su CD-ROM viene fissato in modo tale da controllare il rischio di identificazione sia da possibili individuazioni spontanee che da quelle

[16] http://www.abs.gov.au/Websitedbs/D3110129.NSF/f578250c9c9b9ee1ca256de4002ca08b/95b232c3f39b2022ca256f4a001098bb!OpenDocument!

[17] http://www.abs.gov.au/websitedbs/D3110126.NSF/4fe99d527cb22dbcca256ace00039b49/cb570821d2dcdc02ca256dab0039efb1!OpenDocument

[18] http://www.abs.gov.au/Websitedbs/D3110129.NSF/0/72d92417a0ba71b5ca256d01002c47a4/$FILE/Responsible%20Access%20to%20ABS%20CURFs%20Training%20Manual_Mar05.pdf

dovute a collegamenti tra il file stesso e altre liste in possesso di terzi. Il tipo di file CURF messo a disposizione su CD-ROM è esclusivamente quello base.
In generale una organizzazione può acquistare un file CURF sottoscrivendo un contratto con l'Australian Bureau of Statistics che regola il comportamento che deve essere tenuto dall'organizzazione nell'utilizzo dei dati. L'organizzazione che riceve il file di microdati è tenuta a proteggere le informazioni contenute nel file. I dati devono quindi essere custoditi all'interno di un sistema di computer che assicura l'accesso ad essi solo ed esclusivamente a persone autorizzate dal contratto stipulato con l'Istituto di Statistica. L'ente che ha acquistato il file CURF deve inoltre:

- assicurare che chi fa uso dei dati non tenti di identificare gli individui contenuti nel file;
- garantire che non venga effettuato nessun tentativo di collegamento tra i dati contenuti nel file e dati presi da collezioni esterne per ottenere delle informazioni riservate
- permettere a dipendenti dell'ufficio di statistica di verificare occasionalmente il rispetto da parte degli utenti delle regole sottoscritte nel contratto.

7.2. Accesso in Remoto (Remote Access Data Laboratory – RADL)[19]

I file CURF accessibili attraverso il sistema RADL sono genere file CURF del tipo *base* ed *esteso* per i quali l'ammontare delle perturbazioni e il livello di dettaglio delle variabili rilasciate sono fissati per garantire il controllo del rischio di identificazioni da possibili individuazioni spontanee. Attraverso tale canale non è possibile accedere direttamente ai dati i quali restano all'interno dell'ambiente informatico dell'Istituto. L'utente, attraverso un canale informatico protetto, può far girare i propri programmi di analisi sul file CURF a cui è interessato e l'output ottenuto, nel caso in cui soddisfi una serie di controlli automatici imposti dal sistema RADL, è immediatamente disponibile via e-mail. Nel caso in cui l'utente ne faccia richiesta, l'output che non supera i controlli automatici può essere sottoposto ad un ulteriore controllo di tipo manuale e quindi rilasciato se si ritiene che rispetti le regole di riservatezza.
Come conseguenza del livello di protezione applicata ai file del tipo base, gli output derivanti da elaborazioni fatte sul tali file non sono sottoposti a controlli ne automatici ne manuali.
Anche per questa procedura di accesso è richiesta la sottoscrizione di un contratto che stabilisce una serie di regole da rispettare sia nella fase di elaborazioni (limitazioni sul tipo di istruzioni da sottomettere ai dati) che nell'uso degli output rilasciati.

7.3. Laboratorio Dati elementari (ABS Site Data Laboratory – ABSDL)[20]

I laboratori ABSDL sono predisposti all'interno delle sedi dell'Australian bureau of Statistics. I file CURF messi a disposizioni in tali sedi sono quelli del tipo *speciale,* cioè quelli che mantengono il massimo contenuto informativo compatibile con il rischio di identificazioni spontanee. Attraverso il laboratorio è possibile analizzare anche file del tipo esteso.
All'intero del laboratorio l'intero processo di elaborazione e gli output prodotti sono controllati da personale dell'Istituto. Come per i precedenti canali di diffusione è richiesta la sottoscrizione di un contratto.

---

[19] http://www.abs.gov.au/Websitedbs/D3110129.NSF/0/72d92417a0ba71b5ca256d01002c47a4/$FILE/Responsible%20Access%20to%20ABS%20CURFs%20Training%20Manual_Mar05.pdf

[20] http://www.abs.gov.au/Websitedbs/D3110129.NSF/0/72d92417a0ba71b5ca256d01002c47a4/$FILE/Responsible%20Access%20to%20ABS%20CURFs%20Training%20Manual_Mar05.pdf

8. Sistema Statistico USA

L'Informazione statistica negli Stati uniti non è gestita da una singola agenzia ma ne sono state create diverse ed ognuna di essa si occupa di un particolare settore della vita del paese. Le principali agenzie sono:

- *Bureau of Census (BOC)*- si occupa principalmente della raccolta, dell'analisi e della diffusione dell'informazione statistica derivante dalle indagini censuarie e campionarie sulla popolazione;
- *National Center for Health Statistics (NCHS)* - si occupa del trattamento dell'informazione statistica necessaria per aiutare le azioni e le politiche del paese utili a migliorare la salute della popolazione. Vengono raccolte informazioni sulla nascita, sulla morte e specifiche informazioni mediche sia attraverso delle rilevazioni campionarie che attraverso collaborazioni con specifici centri di ricerca medica.
- *Bureau of Labour Statistics (BLS)* - ha come obiettivo quello di produrre analizzare e diffondere informazioni statistiche sulle condizioni sociali ed economiche del paese in particolare sulle condizioni del mercato lavorativo statunitense.

Inoltre si ricordano le seguenti agenzie: *Bureau of Justice Statistics* (BJS), *Bureau of Transportation Statistics* (BTS), *Division of Science Resources Studies* (SRS), *Economic Research Service* (ERS), *Federal Geographic Data Committee* (FGDC), *National Agricultural Statistics Service* (NASS), *National Biological Service* (NBS), *U.S. Geological Survey* (USGS), *U.S. Geophysical Data Center* (NGDC).

Le Agenzie Federali di Statistica collezionano una grande quantità di informazione che viene poi diffusa sia in forma tabellare che in forma di microdati. Il rilascio di informazione statistica, di qualsiasi tipo, è comunque vincolata al rispetto di una serie di regole che permettono di garantire il diritto alla riservatezza dei rispondenti all'indagine.
In particolare, per il rilascio di microdati, si sono sviluppate una serie di tecniche di protezione perturbative e di canali di diffusione che, combinati opportunamente, permettono, agli utenti che ne fanno richiesta, di disporre di informazioni individuali più o meno dettagliate garantendo comunque la tutela della riservatezza degli intervistati. Analizzeremo in particolare l'approccio seguito dal Bureau of Census e successivamente faremo una panoramica degli approcci seguiti da alcune altre agenzie statunitensi.

8.1. *Bureau of Census* (*BOC*)

Come premesso il Bureau of Census si occupa principalmente di indagini censuarie. Il primo censimento negli Stati Uniti si è svolto nel 1790 direttamente sotto la responsabilità del segretario di Stato[21]. Successivamente, l'espansione demografica e l'incremento dell'informazione statistica raccolta tramite le indagine censuarie ha fatto si che nel 1902 una legge del governo federale ha reso il *Bureau of Census* un'istituzione permanente. I principi di riservatezza che regolano la diffusione di informazione statistica sono quindi l'evoluzione di una lunga esperienza fatta in questi tre secoli di indagini censuarie. Nei primi due secoli l'attenzione rivolta al rispetto della riservatezza dei rispondenti era del tutto nulla o comunque marginale. Successivamente, la complessità delle problematiche analizzate attraverso i censimenti e l'istituzione permanente del Bureau of Census hanno determinato una sempre maggiore attenzione alle problematiche relative alla garanzia della riservatezza dei rispondenti. Attualmente il diritto alla privacy è garantito da una legge federale (Legge n. 13 del Codice degli Stati Uniti) che considera come crimine la diffusione di una qualsiasi informazione che identifichi individui o imprese.

[21] http://www.census.gov/prod/2003pubs/conmono2.pdf 6-2005.

Nello stesso tempo l'agenzia deve tenere conto della responsabilità che ha di diffondere informazione statistica sempre più accurata ed aggiornata, cercando inoltre di soddisfare le sempre maggiori richieste di informazione sotto forma di microdati che vengono dal mondo scientifico. Per fare ciò ha sviluppato una serie di tecniche di protezione che vengono normalmente applicate ai file di microdati che vengono diffusi, ovvero:

- data swapping
- perturbazione casuale dei valore (in particolare per la variabile età)
- ricodifica globale (top-coding)
- soglia sulla densità della popolazione nelle aree geografiche rilasciate

Per quanto riguarda i canali di diffusione messi a disposizione dell'Istituto abbiamo:

- PUMS (Public Use Microdata File)
- Accesso in remoto
- Laboratorio di dati elementari.

8.1.1. Public Use Microdata File (PUMS)

I file PUMS[22] sono file relativi ad un campione di unità familiari e quindi contengono le informazioni sulle caratteristiche di ogni singola famiglia e ogni persona ad essa associata. Nel file vengono escluse tutte le informazioni identificative per assicurare la riservatezza dei rispondenti. Sempre a tale scopo vengono applicate le tecniche di protezione di cui sopra (data swapping perturbazione casuale di alcune variabili ecc) per assicurare la non identificabilità dei rispondenti attraverso le informazioni rilasciate. Durante l'ultimo censimento (2000) a causa dei sempre maggiori sviluppi delle tecnologie informatiche il Bureau of Census ha adottato misure di protezione della riservatezza sempre più stringenti. Allo stesso tempo ha cercato di rispondere alla richiesta di maggiori dettagli geografici attraverso la costruzioni di due tipi di file. Il primo fornisce una vasta gamma di informazioni sociali piuttosto dettagliata (campione all'1% di famiglie) l'altro fornisce una minore quantità di informazioni ma un maggior dettaglio territoriale (campione al 5%).
Il file PUMS 1-percento fornisce il massimo ammontare di informazioni di tipo sociali economiche e familiari disponibili. Non sono stati fissati, a livello nazionale, dei valori soglia per la determinazione delle modalità delle variabili da rilasciare se non per quelle relative alla etnia e all'origine ispanica, soglia fissata ad un valore di 8.000 individui. Per rilasciare tali informazioni è stato necessario un incremento della soglia minima, a livello geografico, superiore a 100.000 individui. Sono state quindi create due entità geografiche ovvero le aree PUMA (Public Use Microdata Area) e super PUMA. Per la costruzione della prima (PUMA) è prevista una soglia minima di 100.000 individui mentre per la seconda (super-PUMA) una soglia minima di 400.000 individui. Le informazioni di tipo geografico diffuse nel file all'1-percento prevedono esclusivamente le aree super-PUMA. L'obiettivo che si è voluto raggiungere con la costruzione di un tale file (1-percento) è quello di costruire un file PUMS che mantenesse le stesse caratteristiche di quelli costruiti per i precedenti censimenti; in particolare per quelli relativi al censimento del 1990.
Per quanto riguarda i file PUMS al 5-percento le informazioni geografiche vengono rilasciate sia a livello di area Puma che a quello delle aree super-PUMA. Per la definizione delle modalità delle variabili da rilasciare è stato fissato un valore soglia nazionale pari a 10.000 individui. I file PUMS sono disponibili sia su CD e DVD oppure possono essere scaricati tramite internet dal sito ufficiale del Bureau of Census.

[22] http://www.census.gov/population/www/cen2000/pums.html 6-2005.

In relazione a file disponibili su internet ricordiamo la banca dati relativa ai cosiddetti file IPUMS (Integrate Public Use Microdata Series)[23]. Tali file sono relativi a 37 collezioni campionarie estratte da più di 50 censimenti federali della popolazione americana e dall'American Community Surveys del 2000-2003. I 37 campioni vengono estratti da ogni indagine censuaria effettuata nel periodo 1850-2000 e vengono protetti attraverso l'applicazione delle diverse tecniche di protezioni precedentemente citate[24]. Il sito dell'IPUMS contiene anche una sezione internazionale in cui sono disponibili collezioni campionarie relative ai censimenti di numerosi paesi.

8.1.2. Accesso in Remoto (America FactFinder)

Per rispondere alle richieste di una maggiore qualità e quantità di informazione statistica il Bureau of Census ha istituito, nell'ottobre del 1998 un sistema di accesso ai dati via Internet: *America FactFinder* (AFF)[25]. Questo sistema, costruito attraverso una collaborazione tra il Bureau of Census e l'IBM Global Service Corp., integra l'esistente sito ufficiale del BOC fornendo al pubblico l'accesso *online* di sempre maggiore informazione. L'accesso ad AFF permette di selezionare la popolazione universo, le aree geografiche e le variabili per la costruzione di tabelle (tabelle non-standard) che non sono previste nel piano di diffusione dei dati da parte del Bureau of Census. In quanto non previste è necessaria una particolare attenzione al rischio di violazione della riservatezza e proprio a tale scopo notevoli sforzi sono stati fatti dal BOC nel determinare regole e tecniche di protezione dei dati.
Attraverso AFF è possibile ottenere solo dati aggregati a partire dai dati relativi ai vari censimenti e altre indagini condotte dal BOC. Per quanto riguarda i dati relativi al censimento del 2000 le informazioni sono fornite solo sotto forma di tabelle. Il BOC diffonde alcune tabelle di tipo base in formato *PDF* che forniscono informazioni demografiche a livello nazionale. Tali tabelle vengono dette prodotti o dati di *primo grado* (*Tier 1 data*).
Vengono chiamati dati di *secondo grado* (*Tier 2* data) le tabelle definite a livello delle cosiddette "r*edistricting area"* ovvero delle aree geografiche ridefinite per tener conto delle necessità amministrative (come le elezioni ecc..). Sono considerate sempre come *dati di secondo grado* più di 300 tabelle predefinite e costruite sulla base dei dati provenienti dalla short form del censimento del 2000 e 800 costruite sulla base della long form. Quando un utente esterno definisce autonomamente delle tabelle attraverso il cosiddetto *Sistema Avanzato di Interrogazione*, le tabelle che ne derivano prendono il nome di prodotti di *terzo grado* (*Tier 3 data*). Proprio il Sistema Avanzato di Interrogazione è considerato il vero e proprio sistema di accesso remoto messo a disposizione dal BOC.
I dati che sono contenuti nel database sul quale avvengono le interrogazioni per la creazione dei prodotti di secondo e terzo grado sono file di microdati precedentemente protetti con tecniche di data swapping e ricodifica globale. Inoltre tutte i prodotti di secondo grado sono in precedenza approvati dal BOC mentre quelli di terzo grado sono diffusi solo se passano attraverso una serie di filtri definiti in base a regole per la limitazione del rischio di identificazione.
Attualmente il Sistema di Interrogazione Avanzato è utilizzato esclusivamente dalle varie agenzie federali di statistica e i vari centri di ricerca federali che hanno una grande esperienza nell'utilizzo di dati censuari. Quindi l'utilizzo di tale canale di accesso ai dati è strettamente legato alle politiche sulla diffusione e protezione dei dati in vigore nel BOC.

---

[23] http://www.ipums.org 6-2005
[24] http://www.ipums.org/usa/codebooks/2000_PUMS_codebook.pdf 6-2005.
[25] http://64.233.183.104/u/census?q=cache:Ua3zN4dcg1YJ:www.census.gov/srd/sdc/AdvancedQuerySystem.pdf+remote+access&hl=en&ie=UTF-8 6-2005.

8.1.3. Laboratorio Dati elementari (Research Data Center – RDC)

Nel 1982 viene creata, all'interno dell'ufficio economico del BOC, una unità operativa di ricerca, ovvero il Census Bureau's Center for Economic Studies (CES)[26], con l'obiettivo di incoraggiare e sostenere la ricerca analitica su microdati di tipo economico e quindi migliorare l'utilità e la qualità dei dati collezionati dal BOC attraverso le regolari indagini campionarie e censuarie. Il CES provvede a mettere a disposizioni dei vari utenti dati confidenziali creando delle strutture protette ovvero i cosiddetti Research Data Center (RDC).
L'accesso da parte di utenti esterni ai microdati è per il BOC un ottimo mezzo per valutare la qualità dei dati da esso stesso collezionati. Infatti bisogna considerare che i dati messi a disposizione degli utenti sono il risultato di un lungo processo di aggiustamento legato ad una serie di decisioni relative a regole di definizione, classificazione, **riservatezza** ecc. che vengono verificate proprio in fase di analisi dei dati da parte di utenti esterni.
I RDC sono locali sicuri gestiti direttamente da personale del BOC. Per la costruzione di tali locali il BOC collabora con accreditate Università e enti di ricerca no-profit. Quindi oltre al RDC localizzato all'interno della sede centrale di Washington esistono diverse sedi sparse in tutti gli Stati Uniti (Boston, Center National Bureau of Economic Research; Los Angeles, University of California; Berkeley, University of Berkeley; Durham, Duke University ecc.). L'utente che vuole accedere ai dati messi a disposizione nei vari RDC deve sottoporre all'approvazione del CES un dettagliato progetto di ricerca. Il CES nella valutazione del progetto verifica la possibilità che esso fornisca dei benefici al BOC, che sia fattibile con i dati contenuti nel laboratorio, sia compatibile con le politiche dell'Istituto e non diffonda informazioni riservate. Nel momento in cui il progetto è accettato al ricercatore viene richiesta la sottoscrizione di un contratto che, per quanto riguarda la diffusione di informazioni riservate, lo rende soggetto alle stesse regole e sanzioni dei dipendenti del BOC. Il ricercatore viene inoltre sottoposto ad una serie di controlli che tra l'altro prevedono anche il deposito delle impronte digitali nell'ufficio del FBI.
I dati messi a disposizione del ricercatore riguardano esclusivamente i dati richiesti dal progetto di ricerca ed esiste una serie di regole da rispettare come la non possibilità di introdurre materiale, di non prelevare dell'output intermedio ecc.
Per minimizzare il rischio di violazione della riservatezza il CES preferisce nettamente progetti di ricerca che prevedano il rilascio di output sotto forma di modelli statistici rispetto a quelli che prevedono il rilascio di tabelle. Il Rilascio di un qualsiasi output è comunque condizionato alla revisione dello stesso da parte del personale del BOC che deve valutare la non diffusione di informazioni riservate.

8.2. *Agenzie Federali di Statistica* (USA)

Le Agenzie Federali di Statistica utilizzano quali canali di accesso ai dati:

- PUMS (Public Use Microdata File)
- Accesso in remoto
- Laboratorio di dati elementari
- Borse di Studio e Programmi di Post dottorato .

In tutti i casi i dati sono preventivamente sottoposti a tecniche di protezione quali:

- data swappig

[26] http://148.129.75.149/CES%20Research%20Report.pdf 6-2005.
http://www.ces.census.gov/ces.php/rdc 6-2005.

- perturbazione casuale dei valore (in particolare per la variabile età)
- ricodifica globale (top-coding)
- soglia sulla densità della popolazione nelle aree geografiche rilasciate

Relativamente ai singoli canali di accesso nelle sezioni successive si descrivono brevemente le politiche adottate da alcune Agenzie Federali.

8.2.1. Public Use Microdata File (PUMS)

Il *National Center for Health Statistics* rilascia più di 500 PUMS che coprono la maggior parte dei campi di ricerca che l'ente persegue. Tali file sono disponibili su CD-Rom o su internet. Sono file protetti con le tecniche di protezione sopra elencate e, sempre per motivi di riservatezza, rilasciati previa accettazione delle restrizioni sull'uso delle informazioni contenute nel file.
Per quanto riguarda il *Bureau of Labour Statistics* vengono prodotti diversi file PUMS sugli individui mentre a causa dell'alto rischio di identificazione sono pochi quelli sulle imprese. Anche il BLS protegge i file PUMS attraverso le tecniche di protezione precedentemente elencate.

8.2.2. Accesso in Remoto

Alcune agenzie, per poter diffondere una maggiore quantità di informazione, rispettando comunque la *privacy* dei rispondenti, hanno sviluppato dei sistemi di accesso in remoto che permettono ai ricercatori di analizzare dati che, per motivi di riservatezza, non sarebbero accessibili tramite file PUMS[27]. I dati analizzabili tramite accesso remoto vengono comunque protetti tramite le normali tecniche di ricodifica, data swapping ecc.
Il sistema di accesso in remoto, attualmente operante presso il NCHS, è un sistema che opera via e-mail. Personale dell'istituto ha costruito dei file di analisi con variabili di comodo strutturalmente uguali ai file di dati reali. In tal modo il ricercatore può produrre i propri programmi di analisi basandosi su tali file. Il sistema si basa sul linguaggio operativo SAS ed è interamente automatizzato. Quindi, una volta inviato il programma via e-mail, è il sistema stesso che acquisisce il programma di analisi, controlla che non contenga operazioni considerate non lecite dal punto di vista della riservatezza (vengono disabilitate funzioni tipo PROC TABULATE PRINT ecc) ed infine lo esegue sui dati reali. Dopo l'esecuzione controlla l'output generato che se non presenta ulteriori problemi viene rilasciato mentre in caso contrario viene sottoposto ad una più approfondita verifica da parte del personale dell'istituto. L'approccio appena descritto si basa quindi sull'esistenza di un data-base, contenente i dati da analizzare, e un sistema analitico di interrogazione.

8.2.3. Laboratorio Dati elementari (Research Data Center – RDC)

La continua richiesta di dati sempre più dettagliati, in particolare con un maggior contenuto informativo a livello territoriale, è stata la spinta che ha portato alla creazione dei vari centri di ricerca dati. Diverse agenzie federali di statistica, così come già visto per il Bureau of Census, hanno creato degli "spazi sicuri" dove, previa autorizzazione, è possibile analizzare dati considerati riservati. Come già notato nel caso dell'accesso in remoto, questi dati, se comunque considerati identificabili vengono protetti con le usuali tecniche di protezione.
Per poter analizzare i dati disponibili presso i vari RDC deve essere presentato un dettagliato progetto di ricerca alla direzione del laboratorio e, se accettato, viene sottoscritto un contratto che regola l'accesso e le procedure di analisi. L'output ottenuto viene rilascio solo previo controllo da parte di personale autorizzato.

---

[27] http://www.bls.gov/ore/pdf/st020380.pdf 6-2005.

A differenza del BOC, sia il laboratorio del NCHS[28] che quello del BLS[29] si trovano esclusivamente nelle sedi principali delle Agenzie in Washington D.C. .

8.2.4. Borse di studio e Programmi di Post-Dottorato

Borse di studio e programmi di post-dottorato forniscono una ulteriore opportunità per i ricercatori che vogliono condurre analisi relative a problemi metodologici ed analitici legati agli obiettivi ed ai programmi svolti dalle varie agenzie di statistica. Attraverso tali strumenti i ricercatori possono condurre le proprie ricerche all'interno delle agenzie utilizzando dati e attrezzature dell'istituto e collaborando con personale interno. Inoltre sono tenuti a rispettare le stesse regole sulla riservatezza a cui sono sottoposti i dipendenti dell'agenzia. I ricercatori che vogliono usufruire di tali opportunità devono presentare un dettagliato progetto di ricerca e informazioni relative alle proprie esperienze formative. I vari progetti vengono poi valutati dagli organi direttivi delle agenzie che propongono tali opportunità

L'American Statistical Association (ASA) in collaborazione con National Scienze of Foundation (NSF) gestisce il programma di borse di studio proposto dal Bureau of Census e dal Baureau of Labour Statistics[30] su metodi di campionamento e sviluppo di metodi per la diffusione dell'informazione statistica. Gestisce inoltre programmi di borse di studio per il National Center of Health Statistics e per il Bureau of Economic Analsis.

Molte agenzie federali prevedono inoltre diversi programmi di post dottorato[31] in svariati campi statistici come record linkage, tutela della riservatezza, campionamento e così via.

[28] http://www.cdc.gov/nchs/r&d/rdc.htm 6-2005.
[29] http://www.bls.gov/bls/blsresda.htm 6-2005.
[30] http://www.census.gov/srd/flyer.pdf 6/2005
[31] http://www.census.gov/hrd/www/jobs/prp.html, 6-2005

9. Accesso remoto – Alcune esperienze

L'accesso remoto è un canale per la diffusione dei dati che ha suscitato notevole interesse da parte del mondo statistico. Diversi sono i progetti internazionali che hanno lo scopo di implementare tale canale e di rendere disponibili dati confidenziali di diversa natura e di diversa origine.
Di seguito descriveremo le esperienze internazionali più significative di implementazioni di tale canale nate dalla collaborazione di Istituti nazionali di statistica e Enti di ricerca.

9.1. *Progetto LIS/LES*

Il progetto LIS[32] (Luxembourg Income Study) (Schouten, 2003) è un progetto di ricerca non-profit nato dalla collaborazione di 25 paesi (tra i quali l'Italia rappresentata dall'Istituto di ricerche sulla popolazione e le politiche sociale – IRPPS e la Banca d'Italia) nel 1983 con l'obiettivo di rendere accessibili microdati provenienti da più paesi per studi comparativi sui redditi familiari.
Il file di base o database del progetto LIS è quindi costituito da una collezione di dati campionari relativi alle indagini sui bilanci delle famiglie che fornisce informazioni demografiche ed economiche sia a livello individuale che familiare. Nel 1994 nasce un progetto del tutto analogo, LES project (Luxembourg Employment Study), che, a differenza del LIS, ha per oggetto microdati campionari relativi alle indagini sulle forze di lavoro provenienti da diversi paesi.
Per entrambi i progetti i microdati vengono dapprima standardizzati e resi confrontabili in modo tale che sia possibile studiare mercati del lavoro e sistemi di reddito tra loro diversi.
I dati contenuti nei due database sono protetti attraverso l'applicazione sia di tecniche di ricodifica che tecniche perturbative[33].
Entrambi i database sono consultabili tramite un sistema di accesso remoto che si basa sul software LISSY appositamente sviluppato. Gli utenti possono accedere al sistema inviando ad un preciso indirizzo, detto mail server, un messaggio di posta elettronica in cui sono contenute le istruzioni da applicare ai dati e le informazioni di identificazione relative all'utente. I programmi che possono essere utilizzati sono SAS, SPSS e STATA. Il mail server è l'unica parte del sistema "visibile".
Il sistema è costituito da diversi componenti software che interagiscono tra loro tramite una o più reti. Le varie componenti quindi ricevono le richieste degli utenti le applicano ai dati e restituiscono i risultati statistici agli utenti tramite posta elettronica. Il ruolo fondamentale è quello svolto dal cosiddetto Post office, ovvero la componente che riceve le richieste. Tale componente ogni 5 secondi riceve le richieste, le analizza dal punto di vista della sicurezza, distribuisce le richieste alla componente che applica i programmi ai dati, invia ai vari utenti le elaborazioni richieste.
Dal punto di vista dei controlli il post office prima verifica la correttezza delle informazioni relative all'utente successivamente esamina la sintassi delle istruzioni di programma richieste per verificare che il tipo di procedura statistica da sottoporre ai dati non violi le regole del progetto relative alla tutela della riservatezza. In particolare ricordiamo che vengono disabilitate le procedure tipo PROC PRINT (SAS), LIST (SPSS e STATA) o FREQUENCIES. I risultati che rispettano tutti i principi di riservatezza fissati dal progetto stesso vengono automaticamente restituiti all'utente. Nel caso il sistema LISSY verifichi una qualche violazione automaticamente rimuove la richiesta dal sistema che viene così esaminata manualmente da personale apposito.

[32] http://www.lisproject.org, 6-2005
[33] http://www.lisproject.org/introduction/faq.htm, 6-2005

9.2. *Progetto PiEP*

Il progetto PiEP[34] (Pay Inequalities and Economic Performance Project) è condotto da una commissione internazionale di ricercatori universitari in stretta collaborazione con Eurostat e alcuni Istituti nazionali di statistica. Il progetto utilizza i microdati relativi all'indagine ESES (European Structure of Earnings Survey) del 1995 di 6 paesi (Belgio, Danimarca, Irlanda, Italia, Spagna, Gran Bretagna). I dati sono conservati nella sede di Eurostat a Lussemburgo e sono consultabili tramite un sistema di accesso in remoto gestito dalla London School of Economics. Il sistema di accesso utilizzato è un adattamento del sistema LISSY descritto nel paragrafo precedente)detto appunto sistema PiEP-LISSY. Questa versione differisce da quella utilizzata per i progetti LIS/LES in quanto prevede dei controlli più restrittivi su comandi o combinazioni di comandi che possono fornire informazioni confidenziali. In particolare vengono disabilitate le procedure per il calcolo delle tabelle di frequenza, la rappresentazione grafica di dati individuali, dei residui e dei valori estremi[35].

10. Conclusioni

L'incremento delle richieste di file di microdati per analisi sempre più specifiche ha creato, per i vari Istituti di statistica, l'esigenza di dover sviluppare metodologie di accesso e tecniche di protezione che, combinate opportunamente, soddisfino al meglio le richieste stesse. Il vincolo imposto dalla tutela della riservatezza dei rispondenti al rilascio e/o all'accesso ai microdati ha comportato la necessità di diversificare i canali attraverso i quali i dati sono resi disponibili.
I tre canali maggiormente diffusi, ovvero, rilascio di file di dati elementari, accesso remoto e laboratori di analisi sono tra loro strettamente complementari ed ognuno di essi presenta vantaggi e limiti. La diversificazione di tali canali è molto legata al tipo di analisi che si vuole condurre ovvero a secondo del tipo di analisi e quindi del dettaglio informativo necessario è possibile pensare di utilizzare un canale piuttosto che un altro.

In particolare i *file di dati individuali* sono facilmente accessibili e risultano essere uno strumento molto flessibile in quanto l'utente è direttamente in possesso dei dati. Ciò comporta però una riduzione del contenuto informativo dovuto all'applicazione da parte degli Istituti di diverse tecniche di protezione per diminuire il rischio di identificazione dei rispondenti. Molti Istituti, (Bureau of Census, Statistics Canada, Australian Bureau of Statistics ecc..) applicano tecniche di protezione che comportano vere e proprie modificazione dei dati (tecniche perturbative).
Uno strumento che viene spesso utilizzato per poter rilasciare maggiori informazioni anche attraverso il rilascio di file di microdati individuali è quello del contratto. Ovvero viene stipulato un vero e proprio contratto tra l'Istituto e l'utente il quale è vincolato a delle regole comportamentali e di utilizzo dei dati permettendo così il rilascio di informazioni più dettagliate.

Negli ultimi anni sempre più Istituti stanno implementando *l'accesso remoto* come canale di accesso ai dati. Una caratteristica importante è la sua flessibilità, in effetti può essere implementato in diversi modi a seconda sia delle esigenze degli Istituti che delle caratteristiche dei dati. Le principali differenze tra le possibili implementazioni sono legate: al tipo di dato che può essere interrogato, al tipo di output che può essere richiesto e infine al tipo di controllo applicato sugli output rilasciati.
Per quanto riguarda il dato interrogato abbiamo due possibili situazioni: un file di dati reali non protetti, come nel caso dell'accesso remoto messo a disposizione da Statistics Canada e da Statistics

[34] http://cep.lse.ac.uk/piep/, 6-2005.
[35] http://cep.lse.ac.uk/piep/papers/Final_Report_V5.pdf, 6-2005.

Denmark, oppure un file di dati protetti come nel caso dei file CURF per l'Australian Bureau of Statistics e dei file messi a disposizione dal Bureau of census su AFF.
Per quanto riguarda invece gli output ottenibili abbiamo casi in cui il sistema di accesso remoto prevede esclusivamente il rilascio di tabelle (AFF); casi in cui gli output rilasciati possono essere esclusivamente il risultato dell'applicazione di modelli statistici (PiEP) o casi in cui sono disponibili entrambi le tipologie di output.
Per quanto riguarda il tipo di controllo che viene effettuato sugli output esso può essere di tipo automatico o manuale. La tipologia del controllo è molto legata al tipo di dati messi a disposizione e all'output ottenibile. Nel caso in cui i dati di base sono dati preventivamente protetti e/o gli output ottenibili sono tabelle il controllo è prevalentemente di tipo automatico, come nel caso di AFF e dell'Australian Bureau of Statistics, ed avviene generalmente attraverso controlli sul tipo di istruzioni che è possibile eseguire. Se invece i dati di base non sono preventivamente protetti, come per Statistics Canada, il controllo è generalmente di tipo manuale.
Per gli utenti, i maggiori vantaggi che offre tale canale sono sicuramente legati sia alla quantità di informazioni fornite che alla facilità di accesso.
Per quanto riguarda l'Istituto, il rendere disponibile i dati tramite tale canale presuppone una notevole mole di lavoro. Infatti, oltre alla parte strettamente informatica relativa alla creazione di un sistema operativo tale da garantire un accesso sicuro tramite internet, esiste una parte strettamente statistica di notevole importanza. E' infatti necessario predisporre datti e metadati di facile consultazione, degli strumenti che permettano la consultazione preventiva dei dati, data set di prova che permettano agli utenti di verificare i propri programmi di anali e l'insieme delle regole che gli output devono soddisfare per garantire la riservatezza dei rispondenti.

I laboratori di dati elementari stanno diventando sempre di più una realtà consolidata per i vari Istituti di statistica. Attraverso tale canale è possibile accedere a dati confidenziali che altrimenti non sarebbero consultabili. Uno degli inconvenienti che deve sostenere l'Istituto è sicuramente legato al costo associato alla realizzazione e alla gestione di tale strutture. Per quanto riguarda gli utenti le critiche rivolte a tale canale sono legate sia ai costi da sostenere per utilizzare e raggiungere le strutture sia al fatto che i dati possono essere analizzati esclusivamente da uno dei software che vengono messi a disposizione all'interno del laboratorio.
Per facilitare gli utenti alcuni Istituti come il BOC e Statistics Canada hanno predisposto delle collaborazioni con alcune Università. Tali collaborazioni (DLI, Statistics Canada) da una parte prevedono un più facile accesso ai dati per chi le sottoscrive dall'altra la possibilità per l'Istituto di realizzare dei laboratori di analisi in locali sicuri all'interno delle Università stesse con l'evidente vantaggio di una offerta maggiormente diffusa sul territorio.

I vari Istituti di statistica stanno facendo notevoli sforzi per aumentare e migliorare l'offerta di microdati. I tre canali sono tra loro strettamente complementari ed ognuno di essi svolge un ruolo particolare e ben definito e si rivolge ad una utenza specifica. Il potenziamento e lo sviluppo di uno dei tre canali non può andare a discapito degli altri e i tre metodi di accesso per ogni singola indagine devono essere sviluppati in maniera collegata.

Riferimenti bibliografici

Andersen, O. (2003). From on-site to remote data access – The revolution of Danish system for access to microdata. *Joint ECE/Eurostat work session on statistical data confidentiality, Luxemburg 7-9 April 2003.*

Istat (2004). Metodologie e tecniche di tutela della riservatezza nel rilascio di informazione statistica. *Metodi e Norme n.20.*

Schouten, B., Cigrang, M.(2003). Remote access system for statistical analisys of microdata. *Statistics and Computing,* 13, 381-389.

Tambay, J.L.,Goldmann, G., Potter, J. (2003). Providing Researcher Access to Data for Analysis at Statistics Canada. *Workshop on Microdata, Stockholm, 21-22 August, 2003*. Disponibile su http://www.micro2122.scb.se/papers.asp

Tranmer, M., Pickles, A., Fieldhouse, E., Elliot, M., Dale, A., Brown,M., Martin, D., Steel, D., Gardiner, C. (2005). The case for small area microdata. *Journal of the Royal Statistical Society, A,*168, part 1, 29-49.