DaWinci/MD: un sistema per data warehouse statistici sul Web

Stefano De Francisci, Leonardo Tininini
Giuseppe Sindoni
ISTAT IASI-CNR

Abstract. Questo lavoro descrive alcuni principi progettuali e un modello innovativo per data warehouse statistici su Web. Questo tipo di sistema si basa su un modello di rappresentazione per tavole statistiche, specificamente pensato per renderle accessibili e navigabili dal Web. Il sistema tiene traccia dello stato della navigazione, la quale è basata su una sequenza di passi di selezione, ciascuno dipendente dalle scelte effettuate nei precedenti. La stessa rappresentazione delle tavole è utilizzata per la memorizzazione dell'informazione sulla disponibilità spazio-temporale nella *meta base dati* del sistema.

1. Introduzione

L'uso crescente del Web come mezzo di diffusione delle statistiche all'utente finale ha recentemente generato molti sforzi di ricerca e sviluppo volti a trovare metodi efficienti per visualizzare e navigare grandi quantità di dati statistici su Internet.

Un *web warehouse* statistico è un sistema di data warehouse [3] specificamente pensato per la pubblicazione di statistiche sul Web, che quindi coniuga l'efficienza e flessibilità dei data warehouse con l'accessibilità e facilità d'uso del Web. Questo significa che i dati sono memorizzati in un data warehouse e sono accessibili dal Web attraverso funzioni di navigazione ipermediale, che permettono all'utente di selezionare e visualizzare dinamicamente e in vari formati i dati desiderati.

Come spiegato nel prosieguo, un web warehouse è particolarmente utile per migliorare quattro tra gli aspetti di qualità dei dati statistici [1]: la puntualità, l'accessibilità, l'interpretabilità e la coerenza. Questo lavoro descrive i principi progettuali e il modello innovativo di un sistema per Web warehouse statistici che è stato sviluppato all'Istat secondo i requisiti di qualità sopra enunciati.

Il sistema presentato è già in esercizio e utilizzato per la diffusione su Internet di varie indagini dell'Istituto. Esso costituisce inoltre una delle componenti del sistema informativo generalizzato di diffusione dei dati statistici dell'ISTAT e nasce dalla sintesi di molteplici esperienze condotte negli ultimi anni all'interno dell'Istituto e volte alla diffusione integrata di dati statistici.

Il lavoro è strutturato come segue. Nelle sezioni seguenti dell'introduzione vengono brevemente illustrati alcuni requisiti e caratteristiche fondamentali del sistema. La sezione 2 descrive alcune analogie e differenze tra concetti statistici e concetti del dominio dei data warehouse. La sezione 3 presenta le funzioni di navigazione del sistema. La sezione 4 descrive gli elementi base dell'architettura del sistema e le loro interazioni. La sezione 5 presenta la struttura del data warehouse, con particolare enfasi sugli aspetti di accessibilità, mentre le conclusioni sono presentate nella sezione 6.

1.1 Web warehouse e qualità dei dati statistici

Come detto in precedenza, mostreremo ora come un web warehouse possa essere particolarmente utile per migliorare quattro tra gli aspetti di qualità dei dati statistici, vale a dire la puntualità, l'accessibilità, l'interpretabilità e la coerenza.

La puntualità delle statistiche si riferisce all'intervallo di tempo tra la loro pertinenza (cioè l'intervallo di tempo al quale si riferiscono) e la loro disponibilità per l'utente finale. Un efficiente web warehouse può aiutare gli statistici a minimizzare il tempo necessario per produrre tavole aggregate e pubblicarle sul Web in modo organizzato. Ne consegue un miglioramento determinante della puntualità, poiché appena i dati aggregati vengono calcolati, essi sono automaticamente disponibili sul Web.

L'accessibilità è la facilità con cui le statistiche vengono messe a disposizione del pubblico dal produttore. La capacità di un web warehouse di pubblicare dati dinamicamente e incrementalmente su Internet lo rende uno strumento molto potente per il miglioramento dell'accessibilità.

L'interpretabilità è la capacità dell'informazione statistica di essere interpretata e utilizzata. Tanta più informazione supplementare viene fornita, maggiore è l'interpretabilità delle statistiche. Come mostrato nel prosieguo, il web warehouse presentato in questo lavoro si basa sui metadati per gestire la navigazione, la selezione, la visualizzazione e l'acquisizione dei dati statistici, permettendone una facile e veloce interpretazione.

La coerenza è la capacità dell'informazione statistica di essere armonizzata nel quadro di un vasto contesto analitico duraturo nel tempo. Questo è esattamente lo scopo per cui sono stati creati i data warehouse. Le tecniche di data warehouse applicate alla pubblicazione delle statistiche costituiscono un'efficiente infrastruttura per il miglioramento della coerenza delle informazioni pubblicate.

1.2 DaWinci/MD e le strategie di sviluppo dell'Istat

Il sistema, nella configurazione illustrata nel presente lavoro, è attualmente in esercizio per la diffusione su Internet dei dati di varie indagini dell'Istituto ([1]) e costituisce una delle componenti del sistema informativo generalizzato di diffusione dei dati statistici dell'ISTAT. Tale sistema, in corso di realizzazione, prevede lo sviluppo di una piattaforma di lavoro integrata per l'analisi e la diffusione in linea dei dati statistici attraverso molteplici *layout*, canali e strumenti di diffusione, operante su sorgenti comuni di dati validati.

Oltre alla componente per la diffusione su Web, il sistema generalizzato è caratterizzato da una pluralità di soluzioni rivolte all'analisi ed alla diffusione interattiva dei dati, tutti da considerare sotto l'ottica delle applicazioni di data warehousing, ma differenziati l'uno dall'altro sulla base della natura dei dati trattati (dati elementari, dati aggregati, tavole statistiche predefinite) e della tipologia di funzioni di accesso e navigazione, nonché, ovviamente, dei canali di diffusione (Intranet ed Internet per le componenti interattive ma anche supporti *off line*, come i CD-Rom da accludere alle pubblicazioni cartacee dei fascicoli provinciali del Censimento della Popolazione).

Finalità principale del sistema generalizzato è quella di consentire alle aree di produzione statistica la costruzione di propri sistemi informativi di diffusione, integrandosi all'interno dei processi di produzione da una parte e sfruttando gli ambienti centralizzati ed i sistemi informativi trasversali dall'altra, per convergere, a livello di Istituto, verso un ambiente integrato e generalizzato di diffusione.

1.3 Inquadramento storico-tecnologico di DaWinci

Il sistema può essere considerato a sua volta come il punto di convergenza di alcune esperienze condotte nel corso degli ultimi anni in ISTAT. Tali esperienze, alcune delle quali tuttora in esercizio, sono rappresentate da:

- Reingegnerizzazione applicativa e tecnologica in ambiente distribuito delle Banche Dati Territoriali (BDT), le quali – sviluppate originariamente in ambiente MVS con tecnologia ADABAS/Natural - hanno per anni costituito il sistema centrale di diffusione a base territoriale dell'Istituto;
- Sistema Informativo sull'handicap, una prima realizzazione di data warehouse basato sul Web, che può essere considerato come progetto "pilota" rispetto a DaWinci/MD;
- Web Warehouse dei dati provvisori dei Censimenti di Popolazione e Abitazioni e di Industria e Servizi del 2001, nonché sue successive evoluzioni per l'indagine sulle acque e il Sistema Informativo sulla Giustizia (DaWinci/PD);

([1]) Al momento è in linea la versione per la diffusione di tutti i dati definitivi del XIV Censimento Generale della Popolazione del 2001; è stata inoltre realizzata una versione del sistema per la diffusione dei dati dell'Indagine sulla Transizione Scuola-Lavoro, in procinto di essere rilasciata; il sistema è stato inoltre recentemente utilizzato per esporre su Web i dati dell'indagine campionaria sui consumi della Bosnia Erzegovina.

Le esperienze menzionate sono confluite nel sistema DaWinci/MD per quanto riguarda la componente di diffusione di dati aggregati su Web. Ad esse si sono aggiunte ulteriori esperienze relative ad altre fasi del ciclo di vita del sistema stesso. Si fa riferimento in particolare a:

- Prototipo del S.I. Territoriale Integrato (SIT-In), attraverso cui è stato possibile gestire le problematiche spazio-temporali del sistema;
- Data Warehouse di Produzione del Censimento della Popolazione, tramite le cui funzioni OLAP[2] è stato possibile procedere all'analisi, aggregazione e interrogazione dei dati elementari nel corso del processo di validazione;
- Sistema di gestione delle procedure ETL[3] per la generazione ed il popolamento di DaWinci/MD a partire da un data warehouse di produzione.

Da ognuno dei sistemi sopra citati - nati per soddisfare specifiche esigenze applicative ma progettati e realizzati proprio in previsione di una soluzione generalizzata - sono state tratte indicazioni ed esperienze utili alla risoluzione di alcuni problemi chiave della diffusione su Web di dati statistici, elementi di conoscenza che sono stati poi riversati ed integrati nel sistema attualmente operante.

1.4 Il modello concettuale di riferimento di DaWinci/MD

Il sistema è basato su un modello di rappresentazione per tavole statistiche specificamente progettato per l'accedere ai dati tramite il Web. Il modello prevede una decomposizione dello spazio informativo in *tavole statistiche semplici*, rappresentate da una coppia di componenti:

- **L'oggetto d'interesse (Cosa)**: la misura (attributo di sommario) della tavola statistica.
- **La classificazione (Come)**: le dimensioni usate per classificare la misura della tavola.

La singola tavola semplice avrà poi molteplici istanziazioni dal punto di vista spazio-temporale:

- **Il tempo (Quando)**: l'istante temporale di riferimento dei dati.
- **Il contesto territoriale (Dove)**: il territorio di riferimento dei dati.

Un aspetto molto rilevante di cui si è dovuto tener conto nella progettazione del sistema è stato come tenere traccia dello stato della navigazione, poiché la navigazione è basata su una sequenza di passi di selezione, ciascuno dipendente dai precedenti. Ciascun passo di selezione contribuisce in modo incrementale e interdipendente alla definizione di ciascuna delle quattro componenti della tavola statistica.

Così un'interrogazione-tipo al sistema consisterà nella specifica di tutte o una parte delle componenti della quadrupla <*t*, *s*, *o*, *c*>, dove t è il riferimento temporale d'interesse, *s* è la combinazione di un dettaglio territoriale *d* e un'area territoriale *a*, *o* è l'oggetto e *c* è una

[2] On Line Analitical Processing, con questo termine si indica una tecnologia software, applicabile anche a database relazionali, che consente l'analisi di grandissime quantità di dati strutturati in viste e gerarchie multidimensionali, producendo risultati il cui significato va oltre quello che si può normalmente ottenere con una query standard.

[3] Gli ETL (Extract, Transform, Load) sono tools informatici che estraggono informazioni dai sistemi transazionali, le trasformano (aggregazione e consolidamento) e le caricano in data warehouse.

combinazione, eventualmente vuota, di classificazioni, per esempio <2001, {"regionale", "isole"}, "Popolazione residente", {"sesso", "stato civile"}>.

La scomposizione delle tavole illustrata sopra è anche alla base del modello di memorizzazione dell'informazione riguardante la disponibilità spazio-temporale delle tavole nella meta base dati del sistema. In tal modo la memorizzazione di una quadrupla <*t*, *d*, *o*, *c*> rappresenterà che la tavola semplice definita dalla coppia <*o*, *c*> è disponibile per l'anno *t* e fino al livello territoriale *d*.

2. Web warehouse e dati statistici

Come già illustrato in [2] esiste una stretta corrispondenza tra basi di dati statistici e data warehouse, che si traduce in alcune regole di modellazione da tenere in considerazione nell'attività di progettazione di un web warehouse efficiente:

- I microdati devono essere strutturati in tabelle di fatti;
- I macrodati devono essere strutturati in cubi e i relativi attributi di sommario alle misure dei cubi;
- Gli attributi di categoria e le relative gerarchie di classificazione devono essere strutturate in dimensioni e livelli dimensionali;
- È necessario prevedere meccanismi efficaci per l'esecuzione di operazioni di roll-up (sommarizzazione), cioè per il passaggio da un livello di aggregazione di maggior dettaglio a uno di minor dettaglio;
- È necessario prevedere meccanismi efficaci per l'esecuzione di operazioni di drill-down, corrispondenti al passaggio da un livello di aggregazione di minor dettaglio a uno di maggior dettaglio.

Questa corrispondenza quasi biunivoca tra concetti statistici e di data warehouse potrebbe far pensare che i data warehouse risolvano completamente tutti i problemi connessi con l'attività di diffusione dei dati statistici. Sfortunatamente, alcune peculiarità che caratterizzano i dati statistici rispetto ai dati operazionali rende necessario estendere le tecniche di data warehouse con modelli e strutture specifici.

La prima peculiarità riguarda le indagini campionarie. Gli aggregati derivanti da micro dati di tipo campionario richiedono normalmente delle verifiche di significatività che non avrebbero alcun senso nei data warehouse convenzionali: tipicamente una collezione di aggregati campionari è considerata significativa sole se, per ciascun gruppo identificato da una combinazione di modalità di classificazione, il numero di unità campionarie supera una soglia predeterminata.

La seconda peculiarità riguarda la privatezza e la *divulgazione secondaria*. La maggior parte dei dati statistici pubblici sono soggetti a leggi nazionali e internazionali che proteggono la privacy dei cittadini. Rispetto agli aggregati, il controllo di divulgazione secondaria implica che

l'organizzazione responsabile delle tavole pubblicate deve essere ragionevolmente sicura che le regole di divulgazione non vengano violate.

Infine, è necessario evitare le *tavole sparse*, cioè tavole statistiche contenenti un'alta percentuale di dati mancanti, dovuta all'uso di combinazioni non significative di classificazioni e/o di modalità di classificazione.

Questi requisiti implicano che i sistemi di data warehouse commerciali non possono essere utilizzati direttamente e senza alcun tipo di adattamento, poiché essi permetterebbero alle applicazioni di navigare arbitrariamente tutte le dimensioni disponibili per un fatto, senza fornire funzioni per controllare la qualità degli aggregati campionari, se la tavola risultante violi le regole di divulgazione e se una data combinazione di classificazioni sia significativa o meno.

Nel resto del lavoro verrà illustrato come questi concetti possono essere applicati nella realizzazione di un sistema finalizzato alla fruizione efficace ed efficiente di tavole statistiche sul Web.

3. Funzioni di navigazione

Questa sezione descrive il meccanismo con cui il sistema permette agli utenti di selezionare e navigare le tavole desiderate attraverso un processo interattivo basato su metadati. Vengono descritte analogie e peculiarità rispetto ai sistemi di data warehouse classici. Si parte da un esempio concreto per fornire una descrizione astratta delle funzionalità.

Un'installazione in esercizio del sistema, con cui vengono pubblicati i dati del XIV Censimento Generale della Popolazione e delle Abitazioni, è accessibile all'indirizzo web http://dawinci.istat.it/MD/.

3.1 I criteri di selezione dei dati

I dati di interesse possono essere scelti secondo quattro parametri: l'oggetto di interesse (ad es. "Popolazione residente"), le classificazioni (ad es. "Sesso", "Età" e "Stato civile"), il territorio di riferimento (ad es. le regioni dell'Italia centrale) e l'anno di riferimento (ad es. 2001). In termini più generali, ciascuna quadrupla di parametri individua un *cubo* di dati, cioè un insieme di valori di una *misura* (l'oggetto) organizzati secondo una struttura multi-dimensionale definita appunto da un insieme di *dimensioni* (le classificazioni, il territorio e il tempo). L'esempio succitato si riferisce a un cubo a cinque dimensioni.

Questo esempio è preso dall'applicazione sviluppata per il Censimento Generale della Popolazione e delle Abitazioni, quindi l'anno di riferimento è sempre il 2001 e la funzione di scelta del parametro temporale è disabilitata. La versione completa del sistema permette agli utenti non solo di scegliere tra diversi istanti di riferimento temporale, ma anche di gestire l'evoluzione temporale delle gerarchie territoriali, come descritto in [4]. Com'è noto infatti, gli elementi di una gerarchia territoriale possono variare nel tempo. Con riferimento alla gerarchia amministrativa, nuove entità territoriali possano essere istituite, destituite o cambiare la posizione nella gerarchia (ad es. comuni che cambiano provincia di appartenenza) per effetto di disposizioni

legislative. Poiché una corretta interpretazione dei dati pubblicati non può prescindere da una corretta ricostruzione di queste dinamiche, un sistema di data wareousing statistico deve essere in grado di gestire queste problematiche, sia dal punto di vista funzionale che, di conseguenza, da quello dei modelli di memorizzazione.

3.1.1 La selezione di oggetti e classificazioni

Un oggetto, un insieme di classificazioni e un contesto territoriale determinano un insieme di *tavole disponibili*, cioè un insieme contenente tutte le tavole aventi come oggetto l'oggetto prescelto o oggetti di maggior dettaglio e come classificazioni le classificazioni scelte o classificazioni di maggior dettaglio, e per le quali sia consentito spingersi fino al dettaglio territoriale richiesto. Questi concetti vengono chiariti meglio nel seguito.

Gli oggetti statistici vengono organizzati in gerarchie: un oggetto può essere genitore di oggetti più specifici o semplicemente un contenitore di altri oggetti. Nel primo caso, tra l'oggetto padre e gli oggetti figli sussiste una relazione di specializzazione: le unità che contribuiscono alle misure relative ai figli sono un sottoinsieme, individuato da una specifica caratteristica, delle unità che contribuiscono alle misure del padre. Nel secondo caso, l'oggetto padre viene definito "artificialmente" per facilitare la navigazione tra gli oggetti disponibili.

Più un oggetto è generico (cioè tanto più esso occupa un posto "alto" nella gerarchia) maggiore è il numero delle tavole statistiche disponibili che ad esso si riferiscono, poiché per quanto detto l'insieme di tavole disponibili per un oggetto è dato da tutte quelle che si riferiscono ad esso più tutte quelle che si riferiscono agli oggetti discendenti dell'oggetto scelto. Ciò permette agli utenti di selezionare un oggetto generico ed ottenere così un gran numero di tavole disponibili, e di posporre ad un passo successivo la scelta della specifica tavola da visualizzare. Questa è una caratteristica molto utile, specialmente per utenti che non hanno una conoscenza approfondita del dominio applicativo, e che non è disponibile in altri sistemi di data warehouse. Ad esempio tali utenti probabilmente non hanno a priori la nozione di "Popolazione residente in famiglia in abitazioni in edificio ad uso abitativo" e delle classificazioni ad essa connesse, almeno non la prima volta che accedono al sistema. Tuttavia, essi avranno quasi certamente la nozione di "Popolazione residente", che è un oggetto a cui corrisponde un vasto insieme di tavole, tra cui quelle corrispondenti a "Popolazione residente in famiglia in abitazioni in edificio ad uso abitativo".

Anche le classificazioni sono organizzate in gerarchie. Una classificazione può essere genitrice di classificazioni più specifiche, o un contenitore di altre classificazioni. Nel sistema del Censimento della Popolazione, la classificazione contenitore "Anagrafiche" raccoglie classificazioni quali "Età" e "Sesso". "Età" è a sua volta madre di varie classificazioni strutturate secondo fasce di età differenti o riferentesi all'età di specifiche categorie di individui. Considerazioni simili a quelle fatte per gli oggetti possono essere fatte a proposito della specificità delle classificazioni e sulla possibilità di scegliere classificazioni generiche per ottenere un insieme più ampio di tavole disponibili.

3.1.2 La selezione territoriale

Ogni tavola è correlata a uno più *contesti territoriali*. Il sistema attualmente in linea per il Censimento, ad esempio, si basa sulle suddivisioni amministrative italiane (dal minimo al massimo livello di dettaglio): Italia, ripartizioni geografiche, regioni, province, comuni e località abitate, mentre è in fase di rilascio un sistema basato sulle suddivisioni amministrative della Bosnia Erzegovina. In generale, è comunque possibile gestire più gerarchie territoriali diverse nella stessa implementazione. È possibile scegliere contemporaneamente l'area di interesse (ad es. Italia Centrale, regione Calabria, provincia di Treviso, etc.) e il livello di dettaglio territoriale (ad es. regionale, provinciale, etc.).

La combinazione di un area e di un livello di dettaglio definisce appunto il contesto territoriale, quindi ad es. le regioni dell'Italia centrale, le province della Calabria o i comuni della provincia di Treviso. Il contesto territoriale, come già detto, è quindi dato dalla coppia $<a, d>$ (area e dettaglio). I due parametri non possono variare indipendentemente, ma le scelte possibili sono dettate dalla specifica struttura gerarchica territoriale. Se quindi chiamiamo d_a il livello "naturale" dell'area scelta e d il livello scelto per il contesto, e se il dettaglio è espresso con numeri interi positivi crescenti al crescere del dettaglio, per ogni contesto deve valere la proprietà $d_a \leq d$. In altre parole il sistema impedisce automaticamente la selezione di contesti territoriali privi di senso come <"regione Calabria", "ripartizionale"> o <"provincia di Rieti", "regionale">.

3.1.3 Aumentare e diminuire i criteri di selezione

Durante la navigazione, ogni scelta successiva del valore di un parametro fa diminuire in generale il numero di valori disponibili degli altri parametri e di tavole disponibili. Quindi ad esempio la scelta di un oggetto provoca:

- la restrizione delle classificazioni disponibili a tutte e sole quelle compatibili con l'oggetto scelto, cioè quelle secondo cui l'oggetto può essere classificato;
- la restrizione dei contesti territoriali disponibili a tutti e soli quelli per cui sono pubblicati dati relativamente all'oggetto scelto;
- la restrizione delle tavole disponibili a tutte e sole quelle relative all'oggetto scelto.

Qualsiasi scelta successiva implica il restringimento delle scelte possibili per i parametri rimanenti. Se quindi ad esempio dopo aver scelto l'oggetto scelgo una classificazione, gli insiemi delle altre classificazioni, dei contesti territoriali e delle tavole disponibili verranno corrispondentemente ristretti.

Questo meccanismo deriva da una scelta progettuale ben precisa: contrariamente a quanto avviene nei data warehouse classici, DaWinci permette di pubblicare solo dati pre-aggregati, per andare incontro ai requisiti di confidenzialità e significatività sopra descritti. Non essendo infatti possibile generare dati mediante calcoli dinamici, non è possibile effettuare scelte che conducano a dati non pubblicabili dal punto di vista della sensibilità o a combinazioni dimensionali corrispondenti a dati inesistenti.

Le combinazioni di massimo dettaglio di classificazioni, oggetti e dettagli territoriali sono determinate preventivamente da parte dei responsabili d'indagine, e quindi definite attraverso

un'apposita fase di progettazione e pianficazione dei contenuti da rendere pubblici. Una volta definite le combinazioni di massimo dettaglio e precalcolati gli aggregati corrispondenti (attraverso tools informatici semiautomatizzati già in fase di consolidamento), la navigazione viene quindi completamente "guidata" dal sistema, caratteristica assai auspicabile per un sistema basato sul Web e che si rivolge ad utenti con abilità informatiche e conoscenza dei dati non note a priori. In pratica, in ciascuna fase della navigazione l'utente sa esattamente quali sono i valori dei parametri già scelti, quali sono ancora disponibili e quante e quali sono le tavole correntemente visualizzabili utilizzando i valori già scelti.

Naturalmente, rimuovere un elemento dall'insieme dei parametri elimina i vincoli imposti dalla presenza di tale elemento e quindi, in generale, comporta un aumento del numero di ulteriori scelte compatibili. Così, per esempio, la rimozione dell'oggetto scelto porterebbe la gerarchia delle classificazioni, dei contesti territoriali e delle tavole disponibili in uno stato che deve essere compatibile soltanto con le classificazioni e il contesto territoriale già scelti e quindi generalmente a liste più numerose e gerarchie più "folte".

Supponiamo quindi di aver scelto "Popolazione residente", <"Sesso", "Stato civile">, <"provinciale", "Calabria">, queste scelte determinano una lista di tavole compatibili e di classificazioni tra cui eventualmente continuare a scegliere per accedere a tavole più dettagliate, cioè tutte quelle che, unitamente a quelle già scelte, danno ancora luogo a tavole semplici abilitate alla diffusione. In questa situazione, la rimozione dalle scelte effettuate dell'oggetto "Popolazione residente" allarga l'insieme delle tavole compatibili, che comprenderà qualsiasi tavola classificata per "Sesso" e "Stato civile" e relativa alle province della Calabria, nonché quello delle classificazioni ulteriormente disponibili, che saranno tutte quelle compatibili con le classificazioni e il contesto territoriale già scelti, a prescindere dall'oggetto di interesse.

3.1.4 La selezione multipla di classificazioni generiche

È possibile continuare a scegliere classificazioni finché ne restano di compatibili con le scelte già effettuate. Il sistema provvede ad avvisare l'utente di questa situazione, che si verifica ogni qual volta viene raggiunto il "limite" dello spazio informativo definito dalle combinazioni permesse di valori di parametri. Si noti che la stessa classificazione generica può essere scelta più di una volta, laddove vi sia almeno una tavola che abbia almeno due classificazioni entrambe discendenti da quella generica. Vale a dire due classificazioni diverse, più specifiche, che siano parte della stessa porzione di gerarchia. Ad esempio si supponga di avere una tavola sulle coppie sposate classificate sia per "Età della moglie" che per "Età del marito" ed un'altra tavola sulle famiglie classificate per "Età della persona di riferimento della famiglia" ed "Età del figlio maggiore". Il sistema permette di selezionare due volte la classificazione generica "Età", e la lista delle tavole disponibili conterrà sia "Coppie sposate per età della moglie ed età del marito" che "Famiglie con figli per età della persona di riferimento della famiglia ed età del figlio maggiore", perché le quattro classificazioni coinvolte sono tutte discendenti di "Età". In questo modo è possibile la selezione generica di tutte le tavole che hanno due classificazioni, diverse, dell'età.

La lista delle tavole disponibili è sempre corredata di metadati, vale a dire una breve descrizione delle classificazioni con le relative modalità. Le pagine HTML contenenti queste descrizioni

vengono costruite dinamicamente, a partire dai metadati memorizzati nella base dati del sistema, durante i passi di scelta dei valori dei parametri. Anche se l'utente ha selezionato delle classificazioni generiche, la lista conterrà per ciascuna tavola la descrizione di dettaglio delle relative classificazioni, in modo che l'utente possa scegliere la tavola che meglio risponde alle sue esigenze informative.

3.2. La pagina di visualizzazione della tavola

Le scelte correnti dei valori dei parametri, ad ogni passo della navigazione sui metadati, determinano una o più tavole che possono essere immediatamente visualizzate. La tavola scelta viene visualizzata in una o più pagine Web, a seconda del numero di classificazioni coinvolte, come illustrato sotto. Tanto più è alto il numero delle classificazioni coinvolte, quanto più cresce la difficoltà nel visualizzare la tavola sul Web. Il problema viene risolto in DaWinci con una tecnica di *paginazione dinamica*.

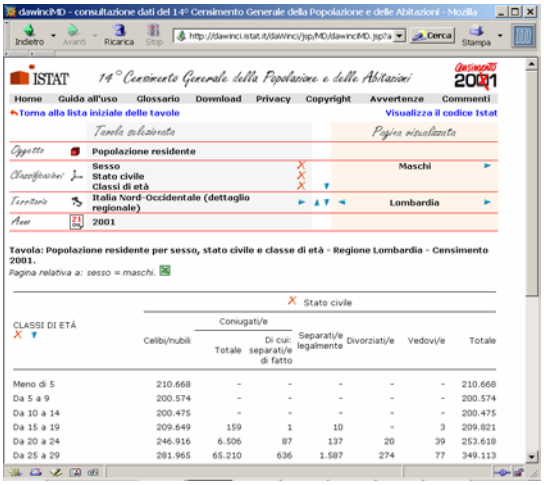

Fig. 1 La pagina di visualizzazione della tavola

La pagina di visualizzazione è divisa in due sezioni: quella superiore è il *pannello di controllo*, i cui collegamenti corrispondono ad altrettante funzioni di navigazione, mentre la sezione inferiore contiene la tavola statistica o una delle pagine in cui la tavola è suddivisa quando sarebbe troppo ampia per poter essere visualizzata in modo agevole all'interno di una sola schermata. Ciascun collegamento sul pannello di controllo corrisponde a una funzione di trasformazione che agisce sulla tavola scelta. Alcuni collegamenti corrispondono a funzioni che visualizzano una pagina differente della tavola stessa, altri corrispondono a funzioni che modificano le scelte dei valori dei parametri, portando implicitamente alla scelta di una tavola differente. Dal pannello di controllo è possibile in particolare:

- eliminare una classificazione, visualizzando così una tavola classificata soltanto secondo le classificazioni rimanenti;
- aumentare o diminuire il dettaglio territoriale, visualizzando così una tavola più o meno dettagliata rispetto al territorio di riferimento;

- variare l'area territoriale, visualizzando così una tavola che si riferisce alla nuova area scelta.

La sezione "Pagina visualizzata" elenca i parametri secondo i quali la tavola è stata suddivisa in pagine. La suddivisione viene effettuata automaticamente dal sistema secondo criteri volti a ottenere un livello soddisfacente di leggibilità e fruibilità concettuale della tavola. La tavola della figura precedente è suddivisa per sesso ed area territoriale (regioni dell'Italia del Nord-Ovest). Per ciascun parametro di paginazione, viene riportata la modalità a cui la pagina corrente si riferisce. Nella figura precedente è visualizzata la pagina relativa ai maschi residenti in Lombardia. È possibile scorrere le modalità delle eventuali classificazioni secondo cui la tavola è paginata, visualizzando così le pagine in successione. Lo stesso discorso vale per le aree territoriali.

Come detto, è possibile rimuovere una classificazione e visualizzare una nuova tavola strutturata secondo tutti i restanti parametri scelti. Questo corrisponde alla scelta di una tavola di minor dettaglio statistico. Dal punto di vista dei data warehouse, questo corrisponde ad effettuare un'operazione di *roll up* sul cubo (tavola) correntemente selezionato. Viene cioè definito un sotto-cubo *(n-1)*-dimensionale a partire da un cubo *n*-dimensionale.

Se sono disponibili ulteriori classificazioni per l'oggetto di interesse corrente, esse appaiono sopra al titolo della tavola. In ogni istante, le classificazioni ulteriormente disponibili sono tutte quelle che, insieme a quelle correntemente selezionate, definiscono una tavola semplice sull'oggetto e per il contesto territoriale correntemente selezionati.

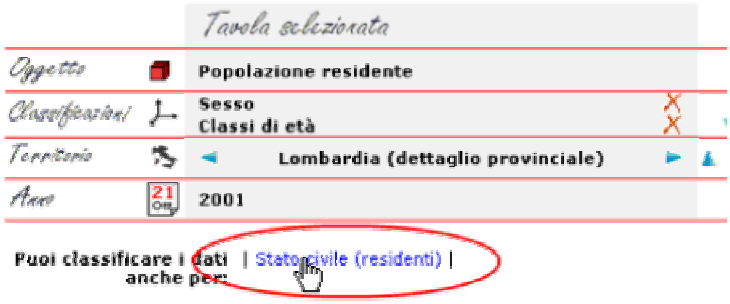

Fig. 2 Aggiunta di una classificazione dalla pagina di visualizzazione

Seguendo il collegamento corrispondente a una classificazione disponibile, viene visualizzata un'altra tavola classificata secondo le classificazioni già scelte più quella nuova, il che corrisponde ad un'operazione di *drill down* secondo la classificazione selezionata. Si passa cioè da un cubo *n*-dimensionale ad un super-cubo *(n+1)*-dimensionale, ovvero a una tavola di maggior dettaglio statistico.

Il livello di dettaglio territoriale della tavola visualizzata può anch'esso essere aumentato o diminuito, ciò risultando in un operazione di drill down o roll up territoriale. In questo caso le dimensioni del cubo non variano, ma varia il livello di una delle dimensioni gerarchiche che

definiscono il cubo. Se il territorio è uno dei parametri di paginazione, per quanto detto sopra, aumentare il dettaglio territoriale significa visualizzare i dati al livello immediatamente inferiore nella gerarchia, limitatamente all'area territoriale della pagina corrente. Per esempio, se sono visualizzati dati dell'Italia nord-occidentale al livello regionale e la pagina si riferisce alla regione Lombardia, aumentare il dettaglio territoriale significa visualizzare i dati della Lombardia a livello provinciale e la prima pagina visualizzata è quella della provincia di Varese (Varese è la prima provincia della Lombardia secondo l'ordine utilizzato nel Sistema Statistico Nazionale). Diminuire il dettaglio territoriale viceversa significa visualizzare i dati al livello immediatamente superiore della gerarchia, limitatamente all'area che nella gerarchia stessa contiene l'area correntemente scelta. Per esempio, se sono visualizzati dati della provincia di Varese a livello comunale, diminuire il dettaglio territoriale significa visualizzare dati della Lombardia a livello provinciale.

Se il territorio non è parametro di paginazione, il dettaglio territoriale può essere aumentato seguendo uno dei collegamenti individuati dai nomi dei territori sulla fiancata della tavola. Viene visualizzata la stessa tavola per il livello sottostante e l'area corrispondente al territorio selezionato seguendo il collegamento. Per esempio, in una tavola che si riferisce alla Lombardia a dettaglio provinciale, seguendo il collegamento "Brescia" si ottiene una tavola di dettaglio comunale che copre la provincia di Brescia. Quando è possibile seguire il collegamento individuato dal totale nazionale, alla fine della fiancata della tavola, viene visualizzata una tavola più dettagliata che si riferisce all'intero territorio italiano. Per esempio, in una tavola a livello ripartizionale che copre tutte le ripartizioni, seguendo il collegamento "Italia" viene visualizzata una tavola di livello regionale che copre tutte le regioni italiane.

Anche il dettaglio di una classificazione gerarchica può essere aumentato o diminuito, risultando in una operazione di drill down o roll up sulla gerarchia corrispondente. Questo corrisponde ancora al passaggio da una tavola di maggiore o minore dettaglio statistico. Per esempio, in una tavola classificata per "Classi di età quinquennali" (0-5, 5-9, …, 95-99, 100 e più) è possibile aumentare il dettaglio della classificazione, ottenendo la stessa tavola in cui il dato relativo a ciascuna fascia d'età[4] viene dettagliato anche per singolo anno (0, 1, …, 89).

3.3. Formato foglio elettronico e mappe tematiche

La tavola correntemente visualizzata può essere memorizzata sul computer dell'utente sotto forma di foglio elettronico. Si noti che ciascun foglio elettronico viene generato dinamicamente, e contestualmente memorizzato anche nel file system del server, solo la prima volta che un utente lo richiede. A ciascun file di foglio elettronico il sistema assegna un nome univoco. Le richieste successive relative alla stessa tavola vengono soddisfatte utilizzando i file già memorizzati, senza bisogno di invocare nuovamente il modulo di generazione dinamica.

È possibile abilitare, per un insieme pre-definito di tavole, la disponibilità di una mappa tematica accessibile attraverso un collegamento Web. La mappa offre una rappresentazione visuale dei

[4] Nell'esempio del Censimento della Popolazione 2001 sono escluse dal dettaglio le ultime tre fasce.

dati, basata sull'associazione dei valori alle corrispondenti aree territoriali, rappresentata utilizzando scale cromatiche o icone a dimensione variabile.

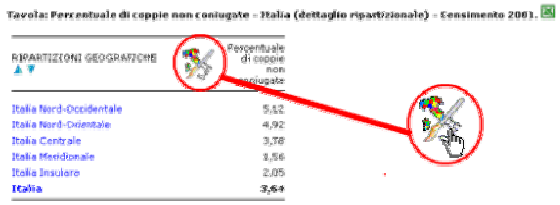

Fig. 3 Collegamento a una mappa tematica

La mappa è generata dinamicamente, a partire dai valori della tavola, da un sistema chiamato *Cartema*.

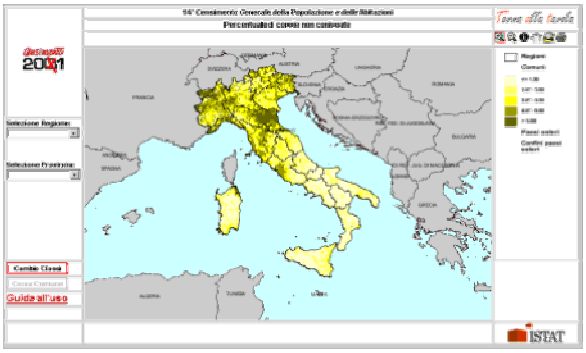

Fig. 4 Visualizzazione di mappe tematiche

Gli utenti possono interagire con la mappa per cambiare il rettangolo di visualizzazione, *zoomare* un'area specifica, abilitare o disabilitare strati geografici e variare gli intervalli numerici per il calcolo delle classi di tematizzazione.

4. L'architettura di sistema e il metadata caching

Questa sezione descrive le componenti architetturali fondamentali del sistema DaWinci/MD, nonché una soluzione più specificamente legata a questioni di ingegneria del software che è però talmente rilevante dal punto di vista delle implicazioni funzionali da essere assimilabile ad una caratteristica architetturale del sistema.

È da notare che, da un punto di vista puramente architetturale, non vi è nessuna differenza tra il sistema DaWinci/MD e il sistema DaWinci utilizzato per la pubblicazione dei dati provvisori dei censimenti (popolazione e industria) e della popolazione legale, nonché della sua ulteriore evoluzione, nota come DaWinci/PD, utilizzata per la pubblicazione dei dati delle indagini sulle acque e delle statistiche giudiziarie.

Le differenze risiedono nel modello di rappresentazione dei dati statistici (quello di DaWinci/MD basato sulla combinazione di oggetti e classificazioni, quello degli altri sistemi più centrato sulla singola tavola statistica) e conseguentemente nel paradigma di navigazione utilizzato per la consultazione (anche se entrambi i sistemi pongono un forte accento sulla esplorazione territoriale del dato statistico), ma l'architettura è rimasta sostanzialmente inalterata, dimostrando eccellenti caratteristiche di scalabilità e affidabilità, nonché capacità di adattarsi alle esigenze e ai modelli di interazione più diversi.

4.1. L'architettura di sistema

L'architettura del sistema DaWinci/MD (ma, come abbiamo già avuto modo di sottolineare, anche degli altri sistemi della "famiglia" DaWinci) è rappresentata nella Figura 5.

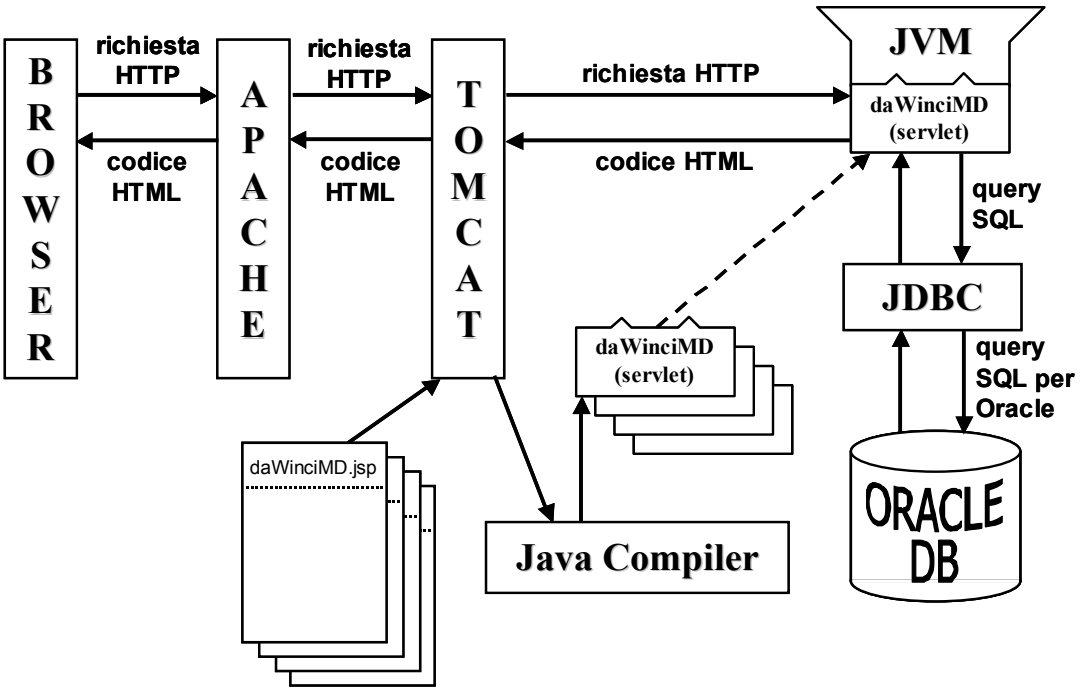

Fig. 5 L'architettura del sistema DaWinci/MD

L'interazione tra la macchina client e il sistema avviene tramite una normale connessione Internet e si esprime attraverso un particolare URL che provoca una richiesta HTTP da parte del browser. Il server Web preposto alla prima gestione delle richieste HTTP da parte degli utenti è attualmente Apache, il più diffuso al mondo, realizzato in tecnologia open-source e disponibile per le architetture hardware/sistema operativo più diverse.

Le richieste HTTP provenienti dai browser degli utenti vengono quindi gestite da Apache che a fronte dell'invocazione di un file JSP (Java Server Page) [5] inoltra la richiesta al server applicativo incaricato di gestire le richieste di questo tipo. Attualmente il server applicativo è

costituito da un altro componente della "famiglia" Apache Jakarta, vale a dire il servlet container Tomcat [6], ma potrebbe essere sostituito da un qualunque server in grado di gestire pagine JSP.

Tomcat utilizza per il proprio funzionamento la Java Virtual Machine (JVM) installata sul server e il relativo compilatore Java. È infatti un particolare tipo di applicazione Java, nota come *servlet*, che si occupa di gestire la richiesta HTTP e di costruirne la relativa risposta. Questa servlet è compilata in formato bytecode e quindi eseguibile su una qualsiasi Java Virtual Machine. A fronte della richiesta di uno dei file JSP del sistema DaWinci/MD, Tomcat verifica se la servlet corrispondente è già disponibile e aggiornata sul server e in caso negativo ne avvia la compilazione producendo la servlet. La servlet viene quindi mandata in esecuzione sulla JVM e Tomcat provvede ad inoltrare la richiesta a sua volta ricevuta dal server Web.

La servlet di DaWinci/MD provvede quindi ad estrarre i dati e metadati necessari alla costruzione della pagina HTML di risposta utilizzando JDBC, lo strato di connettività al database che consente di astrarre dal database specifico utilizzato per la memorizzazione, nella fattispecie un DBMS Oracle. Tramite i dati e metadati estratti dal database la servlet costruisce dinamicamente una pagina HTML che, a seconda della fase di interazione, conterrà gli elementi per la selezione della tavola di interesse (ad esempio la gerarchia delle classificazioni oppure la lista delle tavole compatibili) oppure la tavola statistica vera e propria, corredata di tutti i collegamenti ipertestuali, che consentano all'utente una navigazione di tipo multidimensionale sulla tavola visualizzata. La pagina HTML così costruita risale quindi la cascata applicativa (server applicativo e server Web) raggiungendo infine la macchina client, nella quale il codice HTML viene interpretato e mostrato attraverso la finestra del browser.

4.2. Esportazione dei dati in formato foglio elettronico

A fronte di una richiesta di esportazione della tavola corrente in formato foglio elettronico (Excel), viene invocato un apposito componente del sistema DaWinciMD (V. Fig. 6). Tale componente verifica se il file è già stato richiesto in precedenza e, in caso negativo, genera il file, utilizzando i moduli messi a disposizione da un package Java open source (HSSF) e memorizzandolo stabilmente sul file system accessibile dal server Web. In entrambi i casi la risposta all'utente è costituita da un file HTML contenente un link ipertestuale al file Excel richiesto, che l'utente può quindi facilmente scaricare sulla propria macchina per ulteriori elaborazioni.

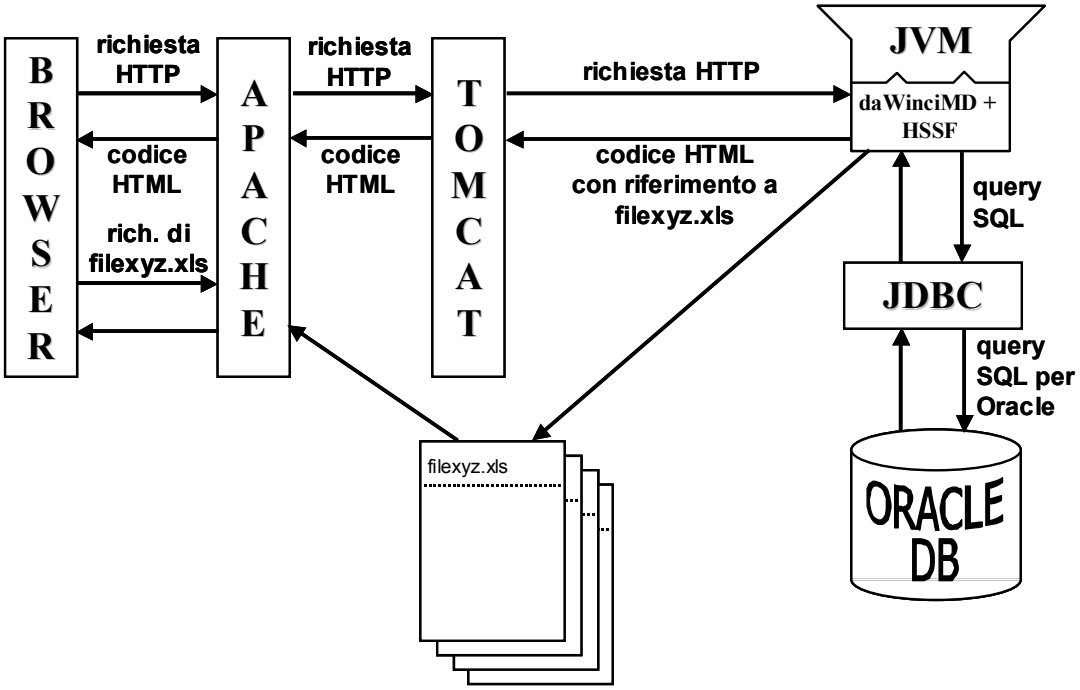

Fig. 6 L'architettura di DaWinciMD e la generazione dinamica con download di tavole in formato Excel

Si noti come questo tipo di architettura consenta un grado di accessibilità molto elevato, in quanto l'utente finale interagisce sempre e soltanto attraverso le pagine HTML visualizzate nel proprio browser. Non è quindi necessaria l'installazione di plug-in addizionali e la navigazione funziona senza problemi perfino con browser "blindati", in cui sia stata disabilitata l'esecuzione di codice Java, Javascript, VBScript, etc.

Si noti altresì che il forte disaccoppiamento tra i vari strati applicativi consente un'elevata flessibilità anche nella scelta delle componenti cruciali del sistema, in particolare server Web, server applicativo e sistema di gestione della base dati (DBMS). Il sistema è infatti in grado di adattarsi con minimo sforzo a server Web e servlet container diversi rispetto a quelli attuali e anche a DBMS diversi (ad esempio Informix o MySQL), richiedendo soltanto interventi marginali su un file di configurazione di DaWinci/MD.

4.3. Tempi di risposta on-line e metadata caching

Uno dei primi problemi che si evidenziò a fronte della sofisticata tecnica di navigazione prevista da DaWinci/MD fu che la costruzione della pagina Web richiedeva un uso molto intenso dei metadati e la necessità di utilizzare query notoriamente non esprimibili in SQL, quali quelle basate sulle relazioni ricorsive di parentela tra i vari oggetti e le classificazioni o su visite in preordine di strutture ad albero (con complesse operazioni di potatura determinate dalle selezioni precedenti) per costruire il layout di selezione di oggetti e classificazioni. Il tutto complicato dalle caratteristiche di protocollo *stateless* (senza memoria) tipiche di HTTP. In altre parole c'erano

molte operazioni complicate da fare e ad ogni nuova pagina bisognava "ricominciare tutto da capo".

Il rischio, puntualmente confermato da alcune iniziali realizzazioni prototipali, era che già soltanto le semplici interrogazioni per l'estrazione ed elaborazione dei metadati richiedessero alcuni secondi, compromettendo seriamente i tempi di risposta, caratteristica irrinunciabile di qualsiasi sistema su Web. La soluzione fu quella di precaricare, elaborare e mantenere stabilmente in memoria centrale una versione locale dei metadati presenti nella base dati, realizzando una sorta di *cache* dei metadati presenti nel sistema.

Questa tecnica di *metadata caching* è stata agevolata dalla possibilità, messa a disposizione dal linguaggio di programmazione Java, di definire oggetti static, noti anche comunemente con il termine di *singleton*, condivisi da tutti gli oggetti di una classe (in questo caso condivisi dalle varie istanzazioni del programma dawinciMD.jsp). Quando una nuova richiesta arriva al sistema abbiamo visto che un'istanza della servlet di DaWinci/MD viene mandata in esecuzione sulla macchina virtuale Java. Il sistema verifica come prima cosa se è presente il singleton contenente una copia pre-elaborata dei metadati e, tramite una semplice query, se tale oggetto è allineato rispetto ai metadati presenti nella base dati. Se è così, come avviene nella quasi totalità dei casi, tutta l'elaborazione sui metadati può avvenire in tempi ridottissimi e senza ulteriori accessi alla base dati.

Gli unici casi in cui si rende necessario il refresh della cache è in caso di fermi del sistema o in corrispondenza del rilascio di nuovi dati, ma l'overhead caratteristico del refresh (alcuni secondi) viene "pagato" solo per la prima richiesta del primo utente che si collega. Peraltro si fa in modo che tale utente sia lo stesso che sta inserendo i nuovi dati, il quale, non appena aggiorna la data di modifica dei metadati, invia contestualmente una richiesta al sistema che provoca il refresh. Si noti anche che il disaccopiamento con i metadati introdotto dal metadata caching consente anche un provvidenziale lasso di tempo (arbitrariamente lungo secondo necessità) tra l'aggiornamento della base di metadati e la sua effettiva visibilità al sistema, lasso di tempo che consente di effettuare verifiche di completezza, coerenza e integrità sui nuovi metadati introdotti, prima che essi vengano effettivamente resi visibili al sistema e quindi all'utenza.

5. Il modello di data warehouse

In questa sezione descriveremo brevemente le componenti fondamentali del modello di data warehouse alla base del sistema di navigazione precedentemente descritto. In particolare analizzeremo il significato delle gerarchie definite tra oggetti e classificazioni, nonché la loro importanza nell'agevolare l'accesso ai dati, anche e soprattutto da parte di utenti poco esperti. Illustreremo inoltre una caratteristica peculiare di DaWinci/MD, vale a dire la definizione di classificazioni con modalità gerarchiche e mostreremo come esse consentano da una parte una fruizione più efficace del dato statistico e dall'altra di aumentare il dettaglio informativo, preservando la significatività e la privatezza dei dati diffusi. Vengono ovviamente ripresi, ed illustrati in maggior dettaglio, anche alcuni dei concetti già esposti nella sezione 3.

5.1. Oggetti e gerarchie di oggetti

Il concetto di *oggetto* nel nostro modello di data warehouse corrisponde solo in parte al classico concetto di *misura* dei data warehouse commerciali, in quanto gli oggetti incorporano anche operazioni di *slicing*[5], derivanti dalle specificità dei data warehouse statistici rispetto a quelli convenzionali. Ad esempio in un data warehouse convenzionale "Popolazione residente" e "Popolazione residente di 6 anni e più" non verrebbero modellate come due misure distinte, mentre in un data warehouse statistico esse costituiscono necessariamente due oggetti distinti, in quanto correlate a combinazioni dimensionali diverse. In particolare, ad esempio, il grado di istruzione non è definito per l'intera popolazione residente (in quanto la relativa domanda del questionario viene posta solo agli individui di 6 anni e più) e quindi sarebbe scorretto correlare tale dimensione all'oggetto "Popolazione residente" che deve essere viceversa correlata all'oggetto più specifico "Popolazione residente di 6 anni e più".

Da un punto di vista generale questa caratteristica consegue dalla presenza di domande filtro nel questionario statistico: la presenza o assenza di determinate dimensioni è legata al valore (risposta dell'intervistato) presente su altre dimensioni. Così la dimensione grado di istruzione risulta valorizzata solo se il rispondente ha un'età maggiore di 5 anni, la dimensione professione risulta valorizzata solo se l'intervistato ha risposto di aver svolto un lavoro, etc. Questa articolazione complessa delle dimensioni di analisi è pressoché assente nei data warehouse convenzionali e ha determinato la differenziazione tra il tradizionale concetto di misura e il concetto di oggetto presente nel modello di DaWinci/MD.

D'altro canto la specializzazione di una misura in più oggetti tende a produrre una proliferazione di oggetti che potrebbe disorientare l'utente finale, soprattutto se inesperto. A tal fine gli oggetti del modello sono stati strutturati in gerarchie, in modo che l'utente abbia comunque la possibilità di scegliere oggetti "generici" o intere categorie di oggetti (di alto livello nella gerarchia). È quindi il sistema che si fa carico di specializzare e combinare le scelte generiche operate dall'utente, facendole corrispondere a tavole statistiche (e quindi ad oggetti specifici) effettivamente presenti nel sistema.

In sostanza un oggetto O2 verrà reso figlio di un oggetto O1 se:

(1) O1 rappresenta un raggruppamento di oggetti più specifici, tra i quali figura in particolare O2 (ad esempio l'oggetto "Popolazione residente" può essere reso figlio dell'oggetto "Popolazione")

(2) Entrambi gli oggetti corrispondono alla medesima misura (stessa funzione di aggregazione), ma O2 è ottenuto applicando la funzione di aggregazione ad un sottoinsieme proprio dell'universo di microdati a partire dai quali si ottiene O1: ne consegue che "Popolazione residente di 6 anni e più" può essere reso figlio di "Popolazione residente".

[5] Con il termine di slicing si intende comunemente un'operazione che agisce su un cubo di dati "affettandolo", evidenziando cioè solo quelle sottosezioni del cubo che soddisfano una o più condizioni del tipo dimensione = (o ≤ o ≥) valore

Si noti che mentre le relazioni gerarchiche di tipo (1) sono introdotte prevalentemente per agevolare l'accesso ai concetti, le relazioni di tipo (2) corrispondono ad effettive proprietà sui dati. Si noti altresì che in molti casi esistono modi diversi, tutti concettualmente corretti, di strutturare in gerarchia un insieme di oggetti (in particolare i raggruppamenti di tipo (1) sono del tutto arbitrari) e che il criterio guida in questi casi deve sempre essere quello di agevolare la navigazione da parte dell'utente finale. In altre parole la strutturazione degli oggetti dovrebbe essere oggetto di un'analisi attenta e dettagliata che privilegi l'ottica dell'utente occasionale, pur tenendo presente la semantica delle relazioni gerarchiche introdotte.

5.2. Variabili, classificazioni e gerarchie di classificazione

I concetti di variabile e classificazione (di variabile) in ambito statistico hanno molte analogie con i concetti di dimensione e livello dimensionale nei data warehouse convenzionali, ma hanno anche alcune caratteristiche distintive di cui si è dovuto tener conto nella definizione del modello di DaWinciMD. In maniera abbastanza informale possiamo dire che una variabile è una proprietà (qualitativa o quantitativa) rilevata sulle unità di osservazione. Quindi, esempi di variabili dell'unità di osservazione "individuo residente" possono essere "professione", "stato civile", "età", etc. Ad ogni variabile possono corrispondere una o più classificazioni di variabile che rappresentano modi diversi di raggruppare i valori che la variabile corrispondente può assumere. Così ad esempio "Età per singolo anno" e "Classi d'età quinquennali" sono possibili esempi di classificazione della variabile "età".

Poiché il modello di DaWinci/MD è orientato alla consultazione dei dati, la classificazione è un concetto chiave per la modellazione della tavola statistica, tuttavia, come vedremo oltre, la variabile, intesa come concetto unificante e generalizzante rispetto alle varie classificazioni, può essere utilizzata molto proficuamente per agevolare il processo di selezione della tavola da parte dell'utente.

Come gli oggetti anche le classificazioni possono (e devono) essere organizzate in gerarchie in modo da agevolare l'accesso ai dati e il legame gerarchico descrive sostanzialmente una relazione di generalizzazione tra concetti. In particolare una classificazione C2 potrà essere resa figlia della classificazione C1 se:

(1) C1 è il nome di una variabile e C2 è una delle sue classificazioni (ad esempio "Classi d'età quinquennali" può essere resa figlia di "età") [6]

(2) C1 rappresenta un raggruppamento di classificazioni più specifiche, tra le quali figura in particolare C2 (ad esempio "età" può essere resa figlia di "variabili anagrafiche")

[6] Si noti che in effetti il nome di una variabile rappresenta nel sistema *una qualsiasi classificazione* della variabile stessa. Quindi ciò che la gerarchia effettivamente esprime è che una "qualsiasi classificazione di età" è una generalizzazione della più specifica "Classi d'età quinquennali". Analogamente nel caso (2) una "qualsiasi classificazione di variabile anagrafica" è una generalizzazione di una "qualsiasi classificazione di età".

(3) C2 è una versione più dettagliata della classificazione C1 (ad esempio "Età per singolo anno" può essere resa figlia di "Classi d'età quinquennali"). Questo caso è del tutto analogo alle tipiche gerarchie dimensionali dei data warehouse convenzionali.

Anche in questo caso, come per gli oggetti, è importante notare che la strutturazione gerarchica può avere un impatto rilevante sulla fruizione del sistema da parte dell'utente, soprattutto per i legami gerarchici di tipo (2).

5.3. Gerarchie e interrogazioni generiche

Cerchiamo di chiarire l'importanza delle gerarchie definite su oggetti e classificazioni al fine di guidare l'utente verso una navigazione fruttuosa sui dati con un semplice esempio ispirato dai dati pubblicati dal 14° Censimento Generale della Popolazione ed Abitazioni. Supponiamo che l'utente sia interessato ad una tavola che fornisce i dati di popolazione residente classificati per età e grado di istruzione.

La tavola statistica effettivamente a disposizione nel sistema incrocia l'oggetto "Popolazione residente di 6 anni e più" con le classificazioni "Classe di età quinquennale da 6 in poi" e "Grado di istruzione (10 modalità)", ma è certamente assai improbabile che l'utente sia consapevole della necessità di considerare una versione specializzata di "Popolazione residente" e di "età" per raggiungere i dati desiderati e comunque è ancora più improbabile che sia in grado di determinare quale sia tale versione specializzata, visto che della variabile età esistono oltre 30 classificazioni diverse utilizzate nell'ambito del Censimento.

Grazie alla possibilità di esprimere la richiesta in termini generici, l'utente specificherà semplicemente l'oggetto "Popolazione residente" incrociato per "età" e "grado di istruzione" (anche di questa variabile esistono diverse classificazioni). Sarà quindi il sistema che, sfruttando il legame gerarchico tra "Popolazione residente" e "Popolazione residente di 6 anni e più", tra "età" e "Classe di età quinquennale da 6 in poi", tra "Grado di istruzione" e "Grado di istruzione (10 modalità)", proporrà all'utente la tavola statistica (o in molti casi una lista di tavole) corrispondente ai dati effettivamente disponibili all'interno del sistema.

L'utente può perfino omettere di specificare l'oggetto di interesse e selezionare soltanto la coppia di variabili "età" e "grado di istruzione", ottenendo dal sistema la combinazione di tutti gli oggetti che sono disponibili nel sistema e che risultano classificati per (versioni specializzate di) età e grado di istruzione.

5.4. Classificazioni con modalità gerarchiche

Nei data warehouse commerciali, sebbene possano esistere delle gerarchie tra i vari livelli dimensionali, si assume normalmente che tali livelli abbiano una struttura essenzialmente piatta, un po' come accade, ad esempio, con la classica gerarchia territoriale amministrativa. Al contrario le classificazioni nelle pubblicazioni statistiche hanno spesso strutture irregolari con subtotali e modalità parzialmente sovrapposte. Si consideri, ad esempio, la seguente classificazione del grado di istruzione:

- LAUREA
- *di cui: con specializzazione e/o dottorato*
- DIPLOMA UNIVERSITARIO O TERZIARIO DI TIPO NON UNIVERSITARIO
- DIPLOMA DI SCUOLA SECONDARIA SUPERIORE
- Maturità liceali
- Altri diplomi di maturità (corso 4-5 anni)
- Diploma scolastico di qualifica
- LICENZA DI SCUOLA MEDIA INFERIORE O DI AVVIAMENTO PROFESSIONALE
- LICENZA DI SCUOLA ELEMENTARE
- ALFABETI PRIVI DI TITOLO DI STUDIO
- ANALFABETI

Si può notare come la modalità "diploma di scuola secondaria superiore" venga ulteriormente specializzata (in modo totale ed esclusivo) dalle tre sottomodalità "maturità liceali", "altri diplomi di maturità (corso 4-5 anni)" e "diploma scolastico di qualifica", e la modalità "laurea" venga dettagliata (ma in modo non totale) nella modalità "con specializzazione e/o dottorato", mentre le altre modalità non vengono ulteriormente dettagliate.

Le ragioni per cui vengono introdotte queste classificazioni dalla struttura articolata ed eterogenea possono essere molteplici, in particolare:

- vincoli di significatività sul dato. Questo fattore è particolarmente rilevante in presenza di indagini campionarie: in presenza di classificazioni dettagliate i dati divengono presto non significativi, per cui è possibile dettagliare solo quelle modalità per le quali il campione "regge".
- vincoli legati alla privacy. In un certo senso simile al precedente: le classificazioni troppo dettagliate tendono a rivelare le caratteristiche dei singoli e si possono dettagliare solo quelle modalità che garantiscano opportune soglie di sicurezza.
- è la realtà stessa in esame a impedire un raffinamento omogeneo per le varie modalità: ad esempio è assolutamente ovvio pensare di dettagliare maggiormente la modalità "diploma di scuola secondaria superiore" o "laurea" distinguendo le varie tipologie, mentre non è affatto ovvio e sensato pensare di dettagliare ulteriormente modalità come "licenza di scuola elementare" o "analfabeti".

DaWinci/MD consente di gestire molto efficacemente classificazioni di variabile con modalità gerarchiche. Il modello del sistema consente di esprimere i legami concettuali esistenti tra le modalità e di generare automaticamente i corrispondenti layout sulle fiancate e testate delle

tavole statistiche. La figura 7 rappresenta ad esempio il layout della succitata classificazione del grado di istruzione realizzata da DaWinci/MD per la testata di una tavola statistica.

Grado di istruzione											
Laurea		Diploma universitario o terziario di tipo non universitario	Diploma di scuola secondaria superiore				Licenza di scuola media inferiore o di avviamento professionale	Licenza di scuola elementare	Alfabeti privi di titolo di studio	Analfabeti	Totale
Totale	Di cui: con specializzazione e/o dottorato		Maturità liceali	Altri diplomi di maturità (corso 4-5 anni)	Diploma scolastico di qualifica	Totale					

Fig. 7 Il layout di testata prodotto da DaWinci/MD per una classificazione con modalità gerarchiche

Le classificazioni con modalità gerarchiche presentano diverse analogie con le cosiddette *gerarchie eterogenee*, analizzate in [7], principalmente dal punto di vista delle implicazioni teoriche.

6. Conclusioni

In questo lavoro sono stati presentati i principali risultati raggiunti in ISTAT per la diffusione su Internet di dati statistici attraverso un sistema di web warehousing – DaWinci/MD - sviluppato nell'ambito del Sistema Generalizzato di Diffusione dell'Istituto. Da una parte efficienza e flessibilità, proprie dei data warehouse, e dall'altra accessibilità e usabilità, proprie dei sistemi sviluppati per il Web, sono stati i principi che, trattati congiuntamente, hanno permesso di costruire uno strumento standard rivolto a coprire gran parte delle esigenze di diffusione su Web dell'Istituto.

Puntando su un paradigma di interazione con l'utente basato sul principio della navigazione ipertestuale, il sistema consente infatti agli utenti di accedere in modo efficiente e molto flessibile a tavole statistiche contenenti dati organizzati secondo aggregazioni a dettaglio variabile, decise in sede di progettazione dei contenuti. I percorsi di navigazione non sono predefiniti, ma guidati dalle scelte correntemente effettuate, in modo da permettere all'utente un'esplorazione *libera ma coerente* dello spazio informativo disponibile. Libera perché vengono messe a disposizione le principali funzionalità di analisi interattiva dei dati in un contesto di navigazione ipertestuale, e coerente perché la struttura e il modello organizzativo del sistema garantiscono la totale consistenza dei percorsi di navigazione garantendo al contempo il pieno soddisfacimento dei vincoli di significatività e confidenzialità del dato.

Destinato a soddisfare sia utenza estemporanea e con modesta abilità informatica, sia specialisti di settore interessati a strumenti di analisi interattiva, DaWinci/MD si è già dimostrato soluzione agevolmente applicabile a varie indagini correnti, sia di tipo censuario che campionario, a base territoriale come a base temporale, coprendo sia il campo delle statistiche sociali sia di quelle economiche.

Ringraziamenti

Gli autori ringraziano Maria Cristina Bedeschi, Francesco Bisceglia, Claudia Cianfarani, Paolo Giacomi, Franco Lodovici per l'insostituibile contributo offerto nello sviluppo del sistema.

Bibliografia

1. Statistics Canada. (2003). Quality Guidelines. Disponibile alla URL: http://www.statcan.ca/english/freepub/12-539-XIE/12-539-XIE03001.pdf
2. Shoshani A. (1997). OLAP and Statistical Databases: Similarities and Differences. Proceedings of the PODS 1997 Conference.
3. Kimball R. (1996). The data warehouse toolkit. John Wiley & Sons.
4. M. Paolucci, G. Sindoni, L. Tininini, S. De Francisci. (2002). Spatio-temporal Information Systems in a Statistical Context. Proceedings of the EDBT 2002 Conference.
5. Java Server Pages technology. (2004) http://java.sun.com/products/jsp/.
6. The Jakarta Site – Apache Tomcat. (2004). http://jakarta.apache.org/tomcat/.
7. Lehner, W. (1998). Modelling Large Scale OLAP Scenarios. Proceedings of the EDBT'98 conference. (pp. 153-167).