La modellazione dei processi nel Sistema Informativo Generalizzato di Diffusione dell'ISTAT

Stefano De Francisci, Massimiliano Renzetti, Giuseppe Sindoni
ISTAT

Leonardo Tininini
IASI-CNR

Abstract. Il sistema informativo generalizzato di diffusione dei dati statistici dell'ISTAT si caratterizza come 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*. Essendo la principale finalità del sistema quella di fornire alle aree di produzione ambienti generalizzati per realizzare in proprio le *data warehouse* di analisi e diffusione sui dati prodotti, risulta centrale, per un utilizzo efficiente del sistema, disporre di un controllo completo dei flussi di attività che ne regolano i diversi impianti. Questo lavoro, oltre a descrivere ad un alto livello di astrazione l'architettura applicativa del sistema, è in particolare dedicato all'illustrazione dei processi aziendali sottostanti, e cioè il complesso dei *cicli di servizio* tramite cui è possibile impiantare le diverse istanze del sistema stesso.

1. Introduzione

Il Sistema Informativo Generalizzato di Diffusione dell'ISTAT è stato concepito nell'ottica di soddisfare progressivamente tutte le esigenze più significative di diffusione elettronica dei dati statistici, nel rispetto delle diversità che caratterizzano l'attività dell'Istituto dal punto di vista della natura dei dati prodotti, sia a livello micro sia a quello macro. L'obiettivo alla base del sistema è stato quello di costruire una piattaforma di lavoro standard che permetta il trattamento e la diffusione dei dati statistici a partire da sorgenti comuni di dati, evitandone il ricollezionamento e la replicazione delle procedure, per la realizzazione dei vari prodotti previsti come *output*. Il sistema è per questo costituito da applicativi generalizzati che coprono sia funzionalità di gestione, quali quelle di estrazione, ricodifica e trasformazione ed aggregazione dei dati statistici, sia funzionalità rivolte agli utenti finali, quali ritrovamento, navigazione e presentazione dei dati in vari formati; esso prevede inoltre la progressiva integrazione con gli altri s.i. di diffusione specialistici, con il portale statistico dell'Istituto e con alcuni dei s.i. centrali che trattano metadati e dati elementari validati.

Le esperienze attuali condotte in Istituto in merito alla diffusione elettronica, il variegato contesto organizzativo ed il differente grado di sviluppo dei sistemi informativi nelle aree di produzione, la continua evoluzione di strumenti tecnologici, la crescita dei fabbisogni

informativi dell'utenza hanno suggerito di procedere non nel senso di un unico s.i. monolitico ed onnicomprensivo, né dal punto di vista applicativo né da quello tecnologico**,** ma nella costruzione di un'offerta multilivello generalizzata ed integrata tramite cui consentire alle aree di produzione di costruire in modo autonomo ma coordinato sistemi informativi di diffusione, tagliati su misura dal punto di vista dei contenuti, ma identici dal punto di vista applicativo e tecnologico.

La componente che costituisce l'ossatura centrale del sistema è costituita da un sistema software (per il quale si è adottato l'acronimo *Istar*) dedicato alla realizzazione dei processi generalizzati di progettazione, costruzione ed utilizzo di data warehouse di analisi e diffusione.

Il sistema attuale è il frutto di una attività che ha visto integrarsi nel tempo vari filoni progettuali, i quali hanno permesso di pervenire ad una unica visione della diffusione dei dati statistici dell'ISTAT, a partire dalle prime esperienze di reingegnerizzazione in ambiente *client/database server* delle banche dati di diffusione centralizzate e, passando per vari stadi intermedi costituiti, ad esempio, da sistemi OLAP per l'analisi dinamica di microdati (*Banca dati sulla mortalità*), da sistemi per la diffusione su Internet di tavole di dati statistici predefinite (*DaWinci* per i primi risultati dei Censimenti della Popolazione e dell'Industria), da sistemi di gestione degli strati semantici necessari alla realizzazione delle funzioni di navigazione utente (*Foxtrot*), fino al sistema attuale che di tali esperienze è la sintesi.

Nel seguito del presente documento, nel descrivere l'architettura del sistema si farà riferimento all'assetto definito a seguito dalle varie fasi di progetto fin qui realizzate[1] da cui il corrente stato del sistema è stato generato. Nella sezione 2 saranno descritte le caratteristiche principali del sistema dal punto di vista architetturale. Nella sezione 3 saranno illustrati i principi dei cicli di servizio previsti per la realizzazione delle varie istanze del sistema, mentre nella sezione 4 si entrerà nel dettaglio delle architetture applicative e del dispiegamento tecnologico di Istar. Nella sezione 5 saranno sintetizzate le principali linee evolutive del sistema. La sezione 6 proporrà alcune brevi riflessioni conclusive.

2. Il sistema informativo generalizzato di diffusione dei dati statistici dell'ISTAT

2.1. Il quadro generale del sistema nel contesto di data warehouse statistici

Costruire un data warehouse statistico è attività che, se non delineata in modo preciso all'interno di una vasta gamma di possibili contesti, rischia di essere affrontata in modo generico e far mettere sullo stesso piano e confondere soluzioni rivolte a realtà sensibilmente differenti l'una dall'altra e caratterizzate da requisiti utente a volte anche in contrasto tra loro. Tale "confusione" può riguardare sia aspetti legati alla natura dei dati trattati (microdati o macrodati) sia la tipologia funzionale delle applicazioni rivolte agli utenti finali

([1]) Le attività per la realizzazione del sistema generalizzato di diffusione, iniziate nel corso del 2003, sono state fin qui articolate in una successione di fasi progettuali: nel corso del 2003, in adempimento alle Direttive del Consiglio è stato prodotto uno studio di fattibilità, caratterizzato anche dallo sviluppo di una versione *prototipale* dedicata alla diffusione dei dati del Censimento della Popolazione del 2001, mentre nel 2004, sempre in adempimento alle Direttive del Consiglio, si è proceduto alla redazione di un documento sulle architetture di massima del sistema visto nella sua globalità e si è proceduto alla realizzazione di una versione dimostrativa del sistema.

(interrogazione o navigazione) sia, infine, lo stesso obiettivo di fondo del sistema di data warehousing (sistemi di supporto decisionale caratterizzati da strumenti di analisi interattiva rivolti ad utenti specialisti o sistemi di *reporting* dinamici per utenti estemporanei). L'avvento delle tecnologie basate su Web e di Internet in particolare ha poi ulteriormente complicato il compito dei progettisti di *data warehouse*, allargandone la fruizione verso classi di utenza non conosciute preventivamente, e comunque fortemente differenziate dal punto di vista delle abilità informatiche e delle necessità informative.
In questo scenario, proporre, per di più, meccanismi generalizzati per la realizzazione di data warehouse e – soprattutto – dare ai principi seguiti la dimensione di linee guida e modelli di riferimento per una organizzazione complessa come un Istituto nazionale di statistica risulta compito quanto mai delicato che deve tener conto dell'effettivo contesto in cui ci si muove, proprio per la particolarità che il trattamento dei dati statistici comporta.

La differenza tra database statistici e data warehouse classici è stata ampiamente trattata nel contesto scientifico [1] ed anche all'interno dello stesso ambito dei data warehouse statistici sono state già affrontate le problematiche originate dall'introduzione del Web [2]. Anche a livello metodologico, le modalità di progettazione e sviluppo di *data warehouse* e *data mart* possono essere basate su solidi impianti progettuali [3], utilizzabili anche in ambito statistico. Quello che resta ancora poco trattato è invece la definizione di un quadro complessivo ed organico che differenzi vari filoni progettuali all'interno di una stessa metodologia, la quale, per essere effettivamente generalizzata e far fronte alle varie esigenze che si possono presentare in realtà applicative complesse, deve essere contraddistinta dall'esistenza di distinti percorsi implementativi, da un adattamento flessibile delle architetture applicative e di dispiegamento ai vari casi, e da un approccio che veda la gestione della fase di analisi e diffusione come *ciclo di vita* a sé stante all'interno di una indagine statistica. Un approccio di progettazione integrata a più livelli è già stato perseguito in ISTAT nell'ambito dell'integrazione dei sistemi informativi a base territoriale [4], specialmente in direzione della armonizzazione delle problematiche afferenti a differenti piani progettuali (organizzativo, concettuale, applicativo e tecnologico) – approccio conosciuto in generale come *passo di derivazione* - ma anche tale approccio non prevedeva una diretta integrazione tra elementi procedurali ed elementi funzionali nella modellazione dei processi del S.I.

Per far fronte a quest'ultimo aspetto, l'architettura generale del S.I. Generalizzato di Diffusione, in particolare del suo motore centrale – Istar – è stata pensata in modo da consentire, da una parte, agli *utenti costruttori* (che rappresenta la classe di utenti abilitati alle funzioni di gestione e impianto delle istanze locali del sistema, assumendo per certi versi un ruolo paragonabile a quello di *designer* dei data warehouse commerciali, secondo la terminologia illustrata più avanti) di *confezionare* istanze autonome del sistema, adattandolo ai propri obiettivi e scegliendo – tra quelle disponibili – le soluzioni applicative più adatte allo scopo prefissato e, dall'altra, agli utenti finali di disporre degli strumenti di interazione con il sistema più appropriati. I due principi su cui tale impianto si basa sono costituiti dai concetti di *ciclo di servizio* da una parte e di *toolkit*[2] dall'altra. Entrambi fanno riferimento alla gestione dei processi cui si demanda la messa in produzione delle varie istanze del sistema ma, mentre nel caso dei *toolkit* si tratta di processi di natura applicativa, il concetto di

([2]) Il termine inglese *toolkit*, non direttamente traducibile in italiano, rappresenta un concetto che, se tradizionalmente viene avvicinato a quello di "cassetta degli attrezzi", significa invece molto spesso un insieme di strumenti e componenti che possono essere utilizzati per costruire un sistema unico e personalizzato secondo le esigenze dell'utilizzatore di questo sistema, ed è sotto questa accezione che è stato utilizzato per rappresentare i prodotti software generalizzati di Istar.
Nel dominio della programmazione dei calcolatori, i *toolkit* (noti anche come "widget toolkits" o "GUI toolkits") sono insiemi di elementi costituenti le interfacce grafiche utente. Essi sono spesso realizzati in forma di librerie, o *framework* applicativi. [v. http://en.wikipedia.org/wiki/Toolkit]

ciclo di servizio rimanda a problematiche legate alla gestione dei processi aziendali[3] (per una trattazione articolata del problema si veda ad es. [5]).

Uno schema generale di processo aziendale è illustrato in figura 1.

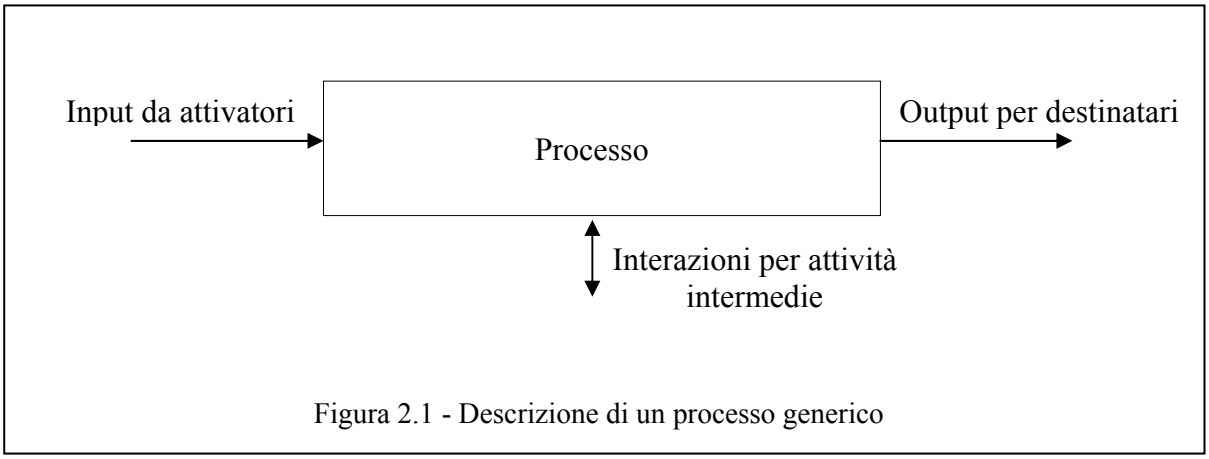

Figura 2.1 - Descrizione di un processo generico

Input ed output, genericamente intesi come prodotti o servizi utilizzati da determinati *clienti* o utilizzatori, sono in questo caso rappresentati dagli ambienti *fonte* e *destinazione* del sistema informativo, mentre il processo aziendale è costituito da una sequenza di passi, lineare nei casi più semplici ma più spesso ciclica, e può richiedere nella sua esecuzione l'interazione con altre fonti di informazioni e di processi, o acquisite direttamente o attinte da ambienti comuni di riferimento.

Se si pensa dunque ad un processo aziendale come l'insieme dei *processi*, collegati tra loro, che devono essere attivati per fornire un certo *output* a partire da *input* definiti e si connotano tali processi sia come procedure informatiche sia come momenti di tipo decisionale, allora tutta l'architettura di Istar può essere intesa come un ambiente di gestione di processi aziendali definiti unicamente in funzione di quale fonte si adotti e di quale ambiente di destinazione si prevede di fornire all'utente finale. In altre parole, la gestione dei flussi di attività del sistema avviene nel rispetto di un approccio che prevede la realizzazione di *processi applicativi guidati da processi aziendali*, nell'accezione sopra delineata.

Spesso però, modellazione dei processi di *business* e modellazione delle applicazioni sono attività condotte non in sufficiente armonia l'una con l'altra. Dare priorità all'analisi delle procedure organizzative, provoca una potenziale riduzione delle soluzioni applicative, che possono essere giudicate – in sede di fase alta di analisi [6] - non fattibili o non efficienti sulla esclusiva base di parametri organizzativi; qualora invece venisse anteposta la logica applicativa al contesto organizzativo, potrebbe verificarsi, nel caso di una cattiva stima dell'impatto sulle modalità di esercizio, un costo gravoso ed inefficiente non solo in termini di interazione con gli utenti e di manutenzione applicativa ma anche del semplice impianto ed avviamento iniziale. Anche utilizzando gli strumenti evoluti di progettazione disponibili per l'uno o per l'altro punto di vista e cioè ponendosi, da una parte, all'interno di logiche di

([3]) I *processi aziendali* (o *business process*) rappresentano funzioni legate all'attività complessiva dell'organizzazione o dell'impresa, quale la produzione di un manufatto, la gestione del personale di una azienda, operazioni connesse all'organizzazione di un servizio di trasporto di merci, l'erogazione di servizi di formazione. Per esempio, per un'organizzazione che si occupa di formazione, sarà necessario definire i processi aziendali per la gestione delle informazioni sui corsi (durata, ambito, propedeuticità, profili professionali interessati, ecc.), sulla loro tempificazione, svolgimento ed esito, sui docenti, sulle richieste ed iscrizioni, sulla composizione delle aule, e così via.

Workflow aziendali[4] [7], pur con strumenti integrati e potenti quali quelli sperimentati in Istat[5] ([8],[9]), dall'altra, adottando metodologie di progettazione orientata agli oggetti in una logica di sviluppo per componenti, non viene garantito che, una volta assemblate le varie componenti del sistema, si raggiungano gli obiettivi fondamentali di efficienza dei processi ed efficacia dei prodotti realizzati.

Pervenire dunque ad una strutturazione formale congiunta di processi aziendali e processi applicativi, ovvero integrare la descrizione dei processi applicativi all'interno di una visione rivolta ai *business process*, tenendo cioè conto sia di aspetti funzionali sia di vincoli organizzativi e procedurali, rappresenta un passo doveroso nella progettazione di sistemi informativi complessi, portando anche ad una indubbia evoluzione ed innovazione dal punto di vista delle metodologie di progettazione. Per far ciò, nel caso del S.I. Generalizzato di Diffusione dell'ISTAT, è stato adottato un approccio progettuale basato sullo sviluppo di pacchetti generalizzati (*toolkit*) all'interno di flussi di attività (*workflow*) definiti sulla base di percorsi incentrati sui prodotti-utente da realizzare.

Infatti, Istar, costruito su una architettura multi-livello e costituito da un complesso di strumenti messi a disposizione di varie classi di utenti, sia in fase di implementazione sia in quella di esercizio, prevede la gestione – assistita quanto più possibile da procedure automatizzate – dei processi applicativi con cui generare i vari livelli di *data warehouse* di dati elementari o aggregati (i prodotti *target*).

Nel contesto del sistema Istar, l'idea del *toolkit* contribuisce allo sforzo di portare a una strutturazione armonizzata i processi aziendali e i processi applicativi, fornendo una struttura concettuale per gli strumenti da utilizzare da parte dei processi applicativi, in un'ottica di realizzazione concreta di un processo aziendale. Più concretamente, il *toolkit* di Istar è un insieme di strumenti applicativi utilizzabili come componenti di sistemi di *workflow* a struttura variabile, tramite cui realizzare ciascuno dei prodotti *target* previsti. Dal punto di vista della logica di organizzazione distinguiamo tra strumenti di definizione e configurazione delle risorse e strumenti di trasformazione delle risorse. In sede di definizione e configurazione delle risorse i componenti del *toolkit* possono essere scelti dall'utente e utilizzati per definire, in modo guidato da metadati sia statistici che operativi, la struttura delle risorse di memorizzazione e gestione dei flussi di processo. In sede di trasformazione delle risorse, altri componenti del *toolkit* possono essere scelti e configurati, sempre in modo guidato dai metadati, in modo da implementare le trasformazioni dei dati attraverso il ciclo di vita definito da ciascuna istanza di *workflow*.

Un contesto siffatto si presta agevolmente al riuso e alla generalizzazione di moduli software preesistenti. Un sistema già sviluppato e correntemente in uso presso l'organizzazione, oppure un sistema a contorno di un'istanza di *workflow*, può infatti essere interfacciato con

---

([4]) Con il termine *Workflow* (letteralmente "flusso di lavoro") si intende l'automazione completa o parziale di un processo di lavoro nell'ambito di un'organizzazione in cui informazioni, documenti e compiti vengono passati da un partecipante ad un altro secondo determinate regole e con un preciso scopo.

([5]) Nell'esperienza ISTAT citata ci si è posto l'obiettivo di effettuare una mappatura dei processi mediante la metodologia *Action Workflow Analysis* (AWA), basata sulla Teoria degli Atti Linguistici, a supporto dell'analisi e della reingegnerizzazione delle attività di produzione statistica. Secondo l'approccio di tale metodologia, flussi di lavoro e processi associati rappresentano in particolare interazioni tra persone, enfatizzando richieste, promesse e assunzioni di impegni reciprocamente attivati. L'attenzione è rivolta quindi ai ruoli svolti da fornitore e cliente di servizi nell'esecuzione dei processi aziendali più che alle azioni ed attività finalizzate alla trasformazione di informazioni, in questo discostandosi dal concetto di *workflow* utilizzato nel presente contesto per descrivere (e gestire) il dipanarsi dei flussi di attività di Istar.

gli altri componenti del *toolkit* scelti, utilizzando varie tecniche di realizzazione dell'interoperabilità tra sistemi software. In particolare, una tecnologia particolarmente in voga è quella degli *Web services*, cioè servizi accessibili per mezzo di messaggi inviati utilizzando protocolli, notazioni e convenzioni sintattico-terminologiche proprie del Web [10]. L'idea di base è di aggiungere a ciascun componente del *toolkit*, visto come mattone fondamentale della costruzione di un processo aziendale, la capacità di "cementarsi" in modo standard, dinamico e riusabile con altri componenti del *toolkit*. Questa facoltà di composizione può essere realizzata *avvolgendo* ciascun componente all'interno di interfacce di ingresso e uscita realizzate con tecnologie Web (tipicamente XML, SOAP[6] e RSS[7]).

2.2. L'architettura multi-layer di Istar

Prima di illustrare come *workflow* e *toolkit* sono stati effettivamente trattati nel sistema, è utile descrivere la sua stratificazione architetturale, in termini di basi di dati, componenti applicative e meccanismi di interazione con gli utenti, riportata in fig. 2.2.

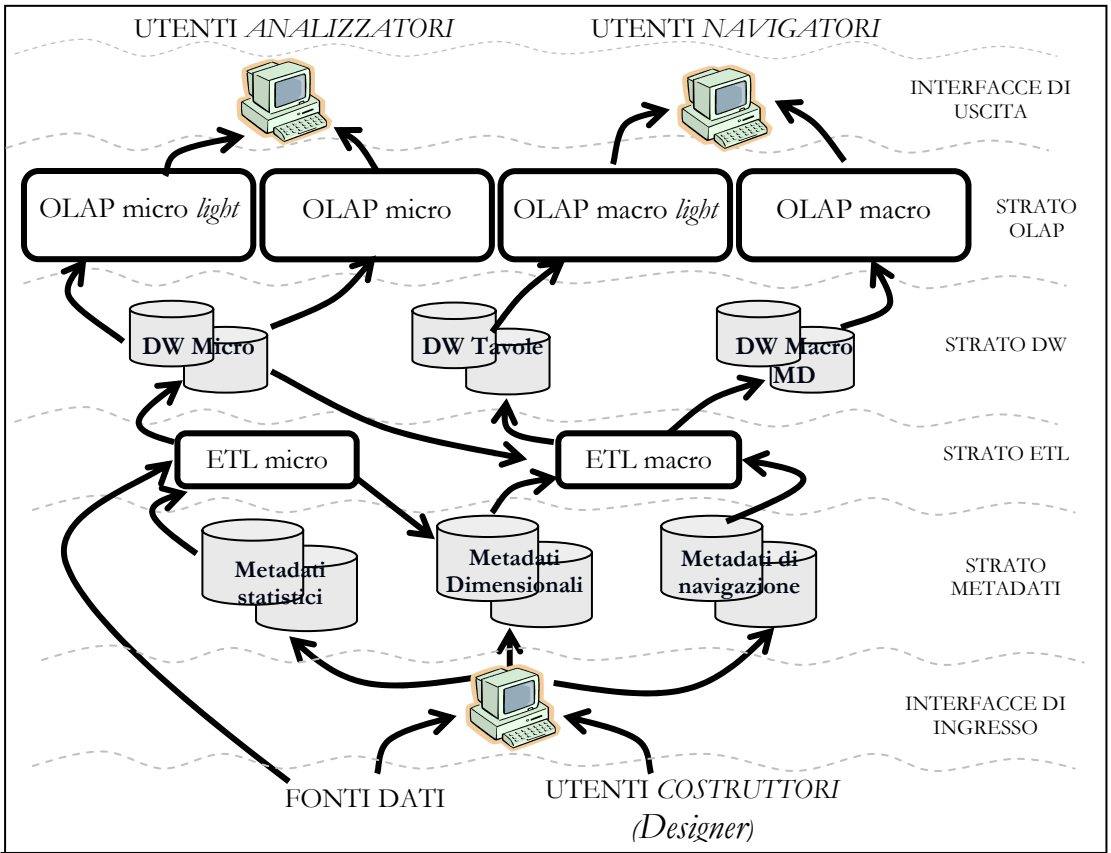

Figura 2.2 – Architettura *multi-layer* di Istar

(6) SOAP sta per *Simple Object Access Protocol* (V. http:// www.w3.org/TR/soap/) è un protocollo leggero per lo scambio di informazioni in un ambiente distribuito e decentrato. Tale scambio di informazioni avviene mediante messaggi codificati in un formato XML.

(7) RSS sta per *Really Simple Syndication* (V. http://blogs.law.harvard.edu/tech/rss) o in alternativa - secondo le versioni dello standard – per *Rich site summary*. È un formato (o più esattamente una famiglia di formati) espresso in Xml, utilizzabile per la diffusione dei contenuti (*content syndication*) sul Web. Attraverso RSS è possibile avere automaticamente una rassegna dei nuovi contenuti immessi in Internet da tutti i siti che, avendolo adottato, abbiano reso disponibili i propri contenuti con questo formato. In tal caso è possibile arrivare direttamente alla pagina che li ospita senza passare per la *home page*. La consultazione è poi facilitata dall'adozione di strumenti chiamati *news aggregator*. Nella forma più semplice RSS contiene una lista (*feed*) di titoli, brevi sommari e link.

Nel quadro sono distinguibili le seguenti componenti:

INTERFACCE DI USCITA: sono costituite da due classi di utenza:

UTENTI ANALIZZATORI cui sono rivolte le funzionalità dinamiche di analisi, aggregazione e *reporting* sui dati elementari. Pur essendo potenzialmente estendibile ad utenza su Internet, si tratta nello specifico di Istar di una tipologia di utenza interna all'Istituto;

UTENTI NAVIGATORI cui sono rivolte le funzionalità dinamiche di interrogazione, ricerca e navigazione su dati aggregati e tavole statistiche. In questo caso si tratta sia di utenti esterni all'Istituto (situazione prevalente) sia interni.

STRATO OLAP: è costituito dalle funzionalità rivolte alle due classi di utenze sopra descritte. Nello specifico si prevede di realizzare due sottoclassi di funzionalità che agiscono su entrambi i livelli di *Data Warehouse*, micro e macro dati (denominate nello schema *completa* e *light*), la stratificazione avverrà sia in funzione delle tecnologie adottate (per la componente micro si prevede di adottare SAS per la versione *light* e *Business Objects* per quella *completa*, per la componente macro invece si tratta di due soluzioni diverse basate sulla evoluzione della tecnologia DaWinci attualmente operante), sia in funzione della ricchezza e complessità applicativa (in genere per la versione *light* si prevede di agire con applicativi maggiormente amichevoli, con un limitato set di funzioni dinamiche e con un minimo impatto esperienziale sulle strutture di produzione);

STRATO DW: è costituito dallo strato di dati *target* (secondo la terminologia adottata in seguito), direttamente accessibili dalle varie classi di utenza attraverso le diverse componenti OLAP.

DW MICRO conterrà microdati statistici provenienti dalle sorgenti di dati validati e strutturati secondo la logica della modellazione *dimensionale statistica* al fine di rendere possibile analisi, aggregazione e reporting in modalità dinamica, sia nella versione *light* sia in quella *completa*.

DW TAVOLE conterrà dati aggregati strutturati in tavole statistiche predefinite, accessibile da utenti esterni e navigabili attraverso interfacce amichevoli. Tipicamente la strutturazione della navigazione (nella versione a livello *light*) sarà orientata e guidata da parametri spazio-temporali e dalle tematiche statistiche di riferimento delle tavole[8].

DW MACRO MD è la Data Warehouse Multi-dimensionale, accessibile dalle funzioni OLAP macro *completa* per la diffusione di dati statistici navigabili secondo il paradigma oggetto-classificazione-spazio-tempo;

STRATO ETL: è costituito dal complesso delle funzionalità di estrazione, trasformazione, aggregazione e caricamento dei dati *trattati* nel sistema che consente, nel suo complesso, il passaggio dal dato nel suo formato sorgente (microdato validato o tavola statistica) al dato da diffondere o su cui effettuare attività di tipo OLAP.

In particolare le funzioni di ETL possono essere suddivise in due principali classi:

ETL1: permette di estrarre dati e metadati dalle sorgenti di microdati validati, integrarli con informazioni aggiuntive costruite interattivamente dagli *utenti costruttori* e, dopo

([8]) Per la definizione delle tematiche attraverso cui effettuare la navigazione verso le tavole statistiche, problematica centrale nell'ambito dei sistemi di accesso alle informazioni su Internet, è ipotizzabile la costruzione di gerarchie di categorie strutturate ad albero, ad esempio secondo la logica classica *broader-narrower*, oppure l'adozione di approcci più complessi, vicini alla semantica, come ad esempio i criteri dizionariali della semantica *a tratti* [11] o quelli basati sui concetti di *lexicon* e *thesauri* della semantica *a istruzioni* [12], criterio preferibilmente rivolto all'analisi testuale, ma ri-orientabile in direzione di una strutturazione categoriale. L'adozione di tali approcci porterebbe ad aumentare il grado di significazione dei contenuti e favorire i meccanismi di gestione della conoscenza.

averne generato le strutture secondo la logica *multi-dimensionale*, alimentare la DW Micro.
ETL2: permette di estrarre dati dalla DW Micro e, attraverso l'integrazione con i metadati di aggregazione e navigazione costruiti tramite la componente *Foxtrot*, di alimentare sia la DW Tavole sia la DW Macro MD;

STRATO METADATI: è costituito dal complesso di informazioni descrittive di natura statistica, informatica, documentale che possono essere meglio classificate come:
MEDATATI STATISTICI (*metastat*), per la rappresentazione dei concetti statistici (oggetti di analisi, variabili, classificazioni, ecc.);
METADATI INFORMATICI (*metadw*), per la rappresentazione dei costrutti relativi alla data warehouse (fatti, misure e dimensioni) e delle regole di ETL;
METADATI DI CONFIGURAZIONE (*metaconf*), per la rappresentazione delle informazioni relative alla generazione e configurazione dei sistemi target;

INTERFACCE DI INGRESSO E GESTIONE: sono costituite da due distinte entità:
ARCHIVIO DEI MICRODATI VALIDATI (***AR.MI.DA.***), che, per *default* viene considerato come la fonte privilegiata da cui estrarre i dati elementari validati (il sistema Istar non ne risulta però vincolato essendo possibile scegliere qualsivoglia fonte di origine)
COMPONENTE INTERATTIVA RIVOLTA AGLI UTENTI COSTRUTTORI (DESIGNER) [9]: Si tratta dei moduli che costituiscono le interfacce per la gestione del sistema Istar e che contengono la logica applicativa tramite cui: rappresentare il dominio statistico di interesse[10]; trasformare i concetti propri del dominio statistico nei corrispondenti concetti propri dei sistemi di data warehouse; gestire tutte le attività del *workflow* che portano alla realizzazione del sistema target; costruire tutto il complesso di metainformazione utilizzata dagli strumenti OLAP di analisi e navigazione.

2.3. La gestione dei processi applicativi: workflow e toolkit

Per quanto detto finora, al fine di definire correttamente il complesso dei *toolkit* necessari al funzionamento del sistema, occorre in sostanza distinguere i quattro elementi costitutivi del *business process* in Istar:

1) gli *ambienti source* (e cioè le fonti da cui sono tratti i dati da diffondere, rappresentati in linea di massima da due livelli: microdati validati e tavole statistiche in qualsivoglia struttura e formato. Ad essi vanno aggiunti i corrispondenti metadati, acquisibili sia interattivamente sia attraverso l'uso di ambienti condivisi);
2) i *sistemi target* (e cioè i prodotti finali costituiti tutti da *data warehouse* in linea, o in ambiente intranet o Internet);
3) i *software* (e cioè le componenti applicative generalizzate, dedicate prevalentemente a due tipologie generali di funzioni: *ETL* e *interazione con gli utenti*);

([9]) Per *UTENTE COSTRUTTORE* si intende, come già detto nel testo, una figura di fatto abilitata a svolgere le funzioni di preparazione, gestione e manipolazione del sistema per predisporre l'impianto delle varie istanze di Istar (prevalentemente attraverso l'organizzazione ed utilizzo di meta-informazioni). In genere, nei sistemi classici di data warehousing, tale figura viene indicata col nome di ***designer***. Il motivo per cui si è voluto adottare il termine "utente" è per porre l'enfasi sul particolare meccanismo di interazione di tale figura col sistema, orientato di fatto al confezionamento del sistema attraverso gli strumenti a disposizione e non a vere e proprie attività di disegno e sviluppo applicativo.
([10]) Con l'espressione *dominio statistico* si intende, nell'ambito del presente documento, l'insieme degli oggetti di analisi (e delle relative variabili e classificazioni) definiti dall'utente costruttore in funzione del sistema target da realizzare.

4) i *workflow* (e cioè i meccanismi attraverso cui si dipana la gestione dei processi applicativi).

GLI AMBIENTI SOURCE:

L'ambiente *source* privilegiato è, come già evidenziato in precedenza, AR.MI.DA., sistema dal quale Istar può estrarre sia i microdati validati, sia i relativi metadati descrittivi dell'indagine considerata.
Tuttavia, sono possibili situazioni nelle quali l'ambiente *source* sia costituito da un sistema sorgente *proprietario* (ovvero un sistema informatico pre-esistente e contenente i dati da elaborare ed eventualmente un insieme di metadati a corredo) o più semplicemente da un set di tavole statistiche già pronte per la diffusione. Si tratta, in questo caso, di scenari nei quali l'area di produzione ha già effettuato una serie di elaborazioni, aggregazioni, controlli, correzioni e validazioni e sarebbe inopportuno obbligare l'utente costruttore a ripetere il lavoro già fatto, ripercorrendo l'intero processo a partire dai microdati validati disponibili in AR.MI.DA.
Tale varietà di ambienti *source* garantisce all'utente costruttore un adeguato livello di flessibilità nell'utilizzo del sistema Istar all'interno dei propri processi produttivi.

I SISTEMI TARGET:

Come già illustrato sopra, i prodotti *target* sono costituiti da quattro ambienti OLAP:
1) Ambiente OLAP su microdati in versione *light* (OLAP-SMOL[11])
2) Ambiente OLAP su microdati in versione *completa* (OLAP-BO)
3) Ambiente OLAP su Tavole statistiche (*DaWinciPD*)
4) Ambiente OLAP su dati aggregati in forma multidimensionale (*DaWinciMD*)

In pratica si tratta sia di sistemi di diffusione propriamente detti, sia di sistemi – a vario livello di complessità – per l'analisi interattiva di dati statistici elementari. I sistemi di diffusione sono basati sulla tecnologia già sviluppata per alcune Data Warehouse dell'ISTAT: DaWinci, nei suoi due principali filoni attualmente disponibili: *DaWinciPD* (per la diffusione di un insieme predefinito di tavole statistiche navigabili rispetto al tempo, al territorio e ad un albero gerarchico di categorie) e *DaWinciMD* (per la diffusione di dati statistici navigabili secondo il paradigma oggetto-classificazione-spazio-tempo [2]). I sistemi di analisi in ambiente intranet sono invece orientati ad attività di tipo OLAP prevalentemente sui microdati, ma che contemplano anche la possibilità di analisi interattiva su dati aggregati.

I SOFTWARE E GLI AMBIENTI GENERALIZZATI:

1) ETL1
2) ETL2
3) FOXTROT.PD
4) FOXTROT.MD
5) FOXTROT.META

Gli strumenti di ETL sono costituiti da motori di estrazione, trasformazione e caricamento che consentono, in complesso, il passaggio dai dati in formato sorgente (microdato validato o tavola statistica) ai dati da diffondere o su cui effettuare attività di tipo OLAP. In particolare, ETL1 è il motore per l'estrazione dei dati e dei metadati dal sistema sorgente di interesse e la

([11]) L'acronimo SMOL sta ad indicare: "Struttura Multidimensionale OLAP Leggera".

creazione e popolamento della data warehouse, mentre ETL2 è il motore per l'estrazione dei dati dalla data warehouse ed il popolamento delle web warehouse dei sistemi di diffusione.
Il complesso di componenti che vanno sotto il nome di *Foxtrot* rappresenta, come già detto, lo strumento principale per la gestione interattiva degli strati semantici del sistema generalizzato. Si tratta di interfacce integrate che includono tutte le componenti software per l'attivazione delle funzionalità di *workflow* e per l'amministrazione della componente metadati del sistema.
I vari moduli di *Foxtrot* contengono la logica applicativa che consente di: rappresentare il dominio statistico di interesse, permettere la trasformazione dei concetti propri del dominio statistico nei corrispondenti concetti propri della letteratura sui sistemi di data warehouse, gestire tutte le attività del *workflow* che portano alla realizzazione del sistema target [12].
In particolare, FOXTROT.META è la componente per la gestione dei metadati di Istar[13]. Si compone a sua volta di un sistema di interfaccia con l'utente interno per la definizione e/o completamento dei metadati statistici e di un sistema per la gestione dei metadati di mapping fra concetti statistici e quelli in ambito data warehouse, articolandosi poi ulteriormente a seconda di quale sistema target si sia prefissato come obiettivo. FOXTROT.PD è la componente per la gestione dei sistemi di diffusione basati su *DaWinciPD*. Infine, FOXTROT.MD è la componente per la gestione dei sistemi di diffusione basati su *DaWinciMD*.

3. I cicli di servizio nel sistema informativo generalizzato di diffusione

Come già messo in evidenza precedentemente, con il termine *workflow* si intende un ciclo di servizio incentrato su una serie di passi che consentono di estrarre informazioni da un sistema sorgente, effettuare elaborazioni su tali informazioni e generare un sistema target di diffusione (per utenti esterni ed interni) o di analisi (riservato ad utenti interni). Per il primo impianto del sistema si prevede di realizzare le seguenti quattro tipologie di *workflow* (distinguibili in funzione dei diversi ambienti source e sistemi target coinvolti):

WORKFLOW 1: dall'archivio dei microdati al sistema di navigazione su web di tavole predefinite basato sulla tecnologia *DaWinciPD*;
WORKFLOW 2: dall'archivio dei microdati al sistema di navigazione multidimensionale su web su dati aggregati basato sulla tecnologia *DaWinciMD*;
WORKFLOW 3: dall'archivio dei microdati al sistema di analisi OLAP *light* su microdati (OLAP-SMOL);
WORKFLOW 4: da tavole statistiche in versione *locale* al sistema di navigazione su web di tavole predefinite basato sulla tecnologia *DaWinciPD*.

Si rende di seguito una descrizione sintetica dei quattro *workflow* previsti, attraverso schede riepilogative dei principali aspetti costituenti cicli e prodotti di lavoro.

([12]) Un tale sistema di amministrazione rende agevole inserimento, cancellazione e aggiornamento dei metadati del sistema, nonché le verifiche di integrità. La sua realizzazione favorisce ad esempio la possibilità di interventi in tempo reale sul rilascio o meno di intere tavole e specifiche classificazioni di riferimento.
([13]) I metadati trattati da *Foxtrot* si caratterizzano secondo due tipologie: la prima riguarda i metadati *semantici*, cioè quelli deputati alla definizione e caratterizzazione del dato diffuso; la seconda tipologia è relativa ai metadati *strumentali*, necessari al reperimento del dato nel database ed alla corretta costruzione e visualizzazione della pagina web. Inoltre, i metadati presenti in *Foxtrot* costituiscono uno degli input al processo di ETL di creazione del macrodato a partire dai microdati validati: tale processo è del tutto generalizzato ed opera ricevendo in input la descrizione delle regole di aggregazione fornite da *Foxtrot*, in termini di metadati operativi.

WORKFLOW 1	
Sistema sorgente	**ARMIDA**
Sistema target	***DAWINCIPD***
	Versione intranet per popolamento e verifica contenuti
	Versione Internet per accesso e navigazione utente finale
Tipologia dati sorgente	**MICRODATI VALIDATI**
Tipologia dati target	**TAVOLE PREDEFINITE**
Sistema metadati	Metadati statistici : estratti da Armida e completati interattivamente con oggetti, classificazioni e relativi incroci Metadati informatici (di popolamento e navigazione) : costruiti tramite mapping da metadati statistici
Struttura Data Warehouse di lavoro	Struttura con schema a stella che costituisce la fonte dati in input al processo di aggregazione e popolamento del sistema target

WORKFLOW 2	
Sistema sorgente	**ARMIDA**
Sistema target	***DAWINCIMD***
	Versione intranet per popolamento e verifica contenuti
	Versione Internet per accesso e navigazione utente finale
Tipologia dati sorgente	**MICRODATI VALIDATI**
Tipologia dati target	**DATI AGGREGATI SECONDO APPROCCIO MULTIDIMENSIONALE**
Sistema metadati	Metadati statistici: estratti da Armida e completati interattivamente Metadati informatici (navigazione utente): oggetti di diffusione e relative classificazioni Metadati informatici (popolamento DW): costruiti tramite *mapping* da metadati statistici
Struttura Data Warehouse di lavoro	Struttura di tipo denormalizzato[14] che costituisce la fonte dati in input al processo di aggregazione e popolamento del sistema

WORKFLOW 3	
Sistema sorgente	**ARMIDA**
Sistema target	**OLAP.SMOL**
Tipologia dati sorgente	**MICRODATI VALIDATI**
Tipologia dati target	**TAVOLE STATISTICHE DINAMICHE**
Sistema metadati	Metadati statistici: estratti da Armida e completati interattivamente con variabili, classificazioni, oggetti di analisi, etc. Metadati informatici (navigazione utente): oggetti di diffusione e relative classificazioni Metadati informatici (popolamento DW): costruiti tramite *mapping* da metadati statistici
Struttura Data Warehouse di lavoro	Struttura con schema a stella che costituisce, nella sostanza, la componente dati del sistema OLAP.SMOL

WORKFLOW 4	
Sistema sorgente	**INSIEME DI TAVOLE IN FORMATO XML**
Sistema target	***DaWinciPD***
	Versione intranet per popolamento e verifica contenuti
	Versione Internet per accesso e navigazione utente finale
Tipologia dati sorgente	**TAVOLE STATISTICHE PREDEFINITE DISPONIBILI LOCALMENTE**
Tipologia dati target	**TAVOLE STATISTICHE PREDEFINITE**
Sistema metadati	Metadati informatici (navigazione utente): tematiche e albero delle aree delle tavole
Struttura Data Warehouse di lavoro	Non esiste struttura di DW di lavoro intermedia

([14]) Tale struttura è in qualche modo derivata dal data warehouse costruito a supporto delle fasi di produzione del Censimento della Popolazione e delle Abitazioni ed in quel contesto indicata come *data warehouse primario*.

Per quanto riguarda le attività necessarie allo sviluppo dei *workflow*, è prevista una articolazione *standard* – generalizzata per tutti i *workflow* sopra definiti ed indipendente dal contesto applicativo di riferimento – basata sui seguenti passi:

1) *Attivazione del workflow*.
2) *Generazione e configurazione del sistema target*.
3) *Estrazione delle informazioni dal sistema sorgente*.
4) *Definizione del dominio statistico*.
5) *Generazione e popolamento della data warehouse*.
6) *Popolamento del sistema target*.
7) *Esercizio del sistema target*.

Tale articolazione del generico ciclo di servizio si concretizza poi in modo specifico in funzione della tipologia di *workflow* preso in considerazione, come risulterà chiaro fra poco. Ciò che è importante mettere subito in evidenza, tuttavia, è che tale differenziazione non ha alcun impatto sulla fruibilità del sistema da parte dell'utente costruttore, il quale viene automaticamente instradato sul percorso corretto dallo stesso sistema Istar, coerentemente con l'ambiente *source* disponibile e con il sistema *target* prescelto.

3.1. Workflow 1: dai microdati al sistema DaWinciPD

In fig.3.1 viene riportato, tramite *diagramma delle attività UML*, lo schema di processo relativo al *workflow* 1, ovvero lo scenario in cui il sistema sorgente fornisce i microdati validati (ed un insieme di metadati a corredo) ed il target da realizzare consiste in un sistema di diffusione basato sulla tecnologia *DaWinciPD*. Come risulta anche dall'illustrazione, in questo caso il *workflow* appena descritto nelle sue linee generali si dipana secondo quanto di seguito descritto:

0. *Attivazione del workflow*. L'utente costruttore intende realizzare un nuovo sistema di diffusione di tipo *DaWinciPD*. A tale scopo, entra in contatto con lo specialista Istar[15], il quale autorizzerà l'utente costruttore all'ingresso nel sistema informatico Istar, secondo un opportuno profilo di accesso.
1. *Generazione e configurazione del sistema target*. Lo specialista Istar predispone le necessarie infrastrutture hardware e software che ospiteranno il sistema target, dopodiché provvede alla sua installazione e configurazione. Al termine di tali operazioni, l'utente costruttore ha a disposizione due ambienti fisici distinti: un sistema di diffusione *DaWinciPD* in ambiente intranet (accessibile tramite meccanismo basato su username/password, da utilizzare per il popolamento e la verifica dei contenuti informativi da diffondere) ed uno esposto su Internet (accessibile a tutti e sul quale saranno di volta in volta caricati esclusivamente i contenuti informativi pubblicati). Entrambi i sistemi di diffusione forniti all'utente costruttore sono inizialmente privi di contenuti informativi statistici. Parallelamente a tali attività di competenza dello

([15]) Con il termine *Specialista Istar* si è voluto indicare nel testo un complesso di figure polivalenti, destinate a svolgere vari poli di attività e di conseguenza orientate a ricoprire molteplici ruoli nei *workflow*. Volendo raffinare il termine, può essere ad esempio individuata la figura di amministratore, quella di supporto statistico, tecnologico od applicativo, ma anche l'*esperto* delle varie declinazioni del sistema che guida gli utenti verso corrette scelte iniziali, e così via. Nella descrizione dei *workflow* sarà sempre utilizzato il termine Specialista Istar al posto delle figure più specifiche proprio per porre l'enfasi sul concetto di *pool* di figure di supporto agli utilizzatori nella costruzione delle varie istanze del sistema..

specialista Istar, l'utente costruttore può già accedere al sistema Istar per compiere le attività di cui ai successivi passi 2 e 3.

2. *Estrazione delle informazioni dal sistema sorgente*. Tramite il motore ETL1 di Istar, l'utente costruttore attiva un processo di estrazione automatica di tutti i metadati statistici disponibili nell'ambiente source, metadati che vengono memorizzati nel database del sistema Istar.
3. *Definizione del dominio statistico*. L'utente costruttore utilizza Istar per visualizzare i metadati estratti dall'ambiente source nel corso del passo 2 e completare eventualmente il proprio dominio statistico, definendo nuove variabili, classificazioni, oggetti di analisi, etc. A fronte dei metadati costituenti il dominio statistico, il sistema Istar attua una fase di mapping automatico che determina l'insieme dei metadati informatici che saranno memorizzati in *database metadati* ed utilizzati nel successivo passo 4 per la generazione ed il caricamento della data warehouse, attività propedeutiche al popolamento del sistema target *DaWinciPD*.
4. *Generazione e popolamento della data warehouse*. Una volta completate le attività *generazione target* e *definizione dominio*, l'utente costruttore può attivare nuovamente il motore ETL1 di Istar che, in funzione dei metadati statistici ed informatici contenuti nel database dei metadati, provvede a generare la data warehouse e a popolarla con i dati estratti dal sistema source ed elaborati secondo opportune regole di trasformazione.
5. *Popolamento del sistema target*. L'utente costruttore, tramite il motore ETL2 ed un'opportuna interfaccia utente del sistema Istar, inserisce nel proprio sistema target tavole statistiche e documentazione a corredo, arricchendone così il contenuto informativo statistico. Tali informazioni vengono via via verificate e validate dallo stesso utente costruttore (tramite il sistema *DaWinciPD* disponibile in ambiente intranet) e reso disponibile alla comunità Internet (utilizzando opportune funzionalità di pubblicazione rese disponibili dal sistema Istar).
6. *Esercizio del sistema target*. L'utente esterno accede via Internet al sistema target *DaWinciPD* così realizzato e fruisce dei contenuti informativi statistici in esso presenti.

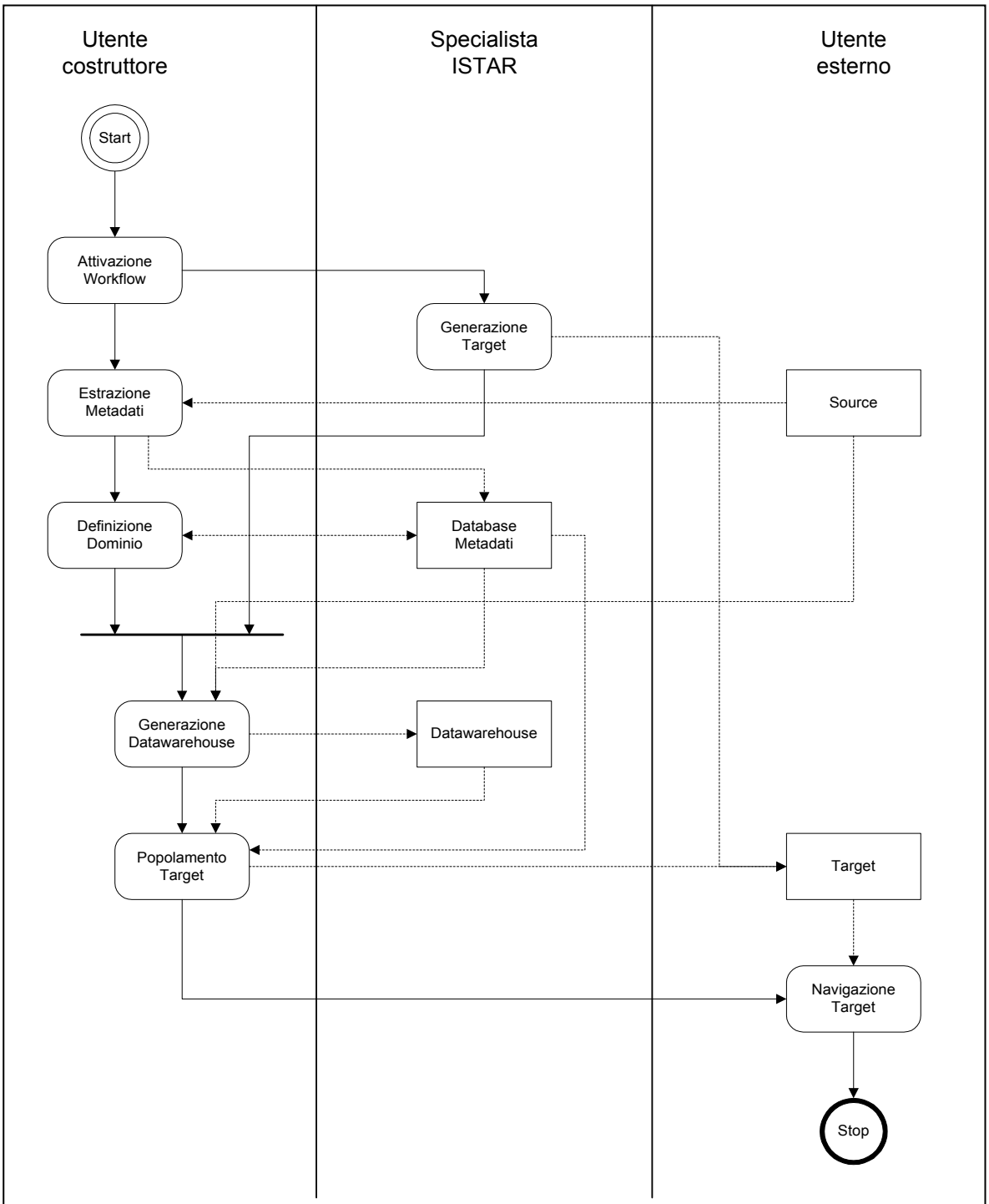

Figura 3.1 – Workflow 1 (dai microdati a *DaWinciPD*) e
Workflow 2 (dai microdati a *DaWinciMD*)

3.2. Workflow 2: dai microdati al sistema DaWinciMD

Come risulta evidente dalla fig.3.1, il *workflow* di tipo 2 presenta una struttura del tutto analoga a quella appena descritta a proposito del *workflow* di tipo 1, con l'unica differenza del terminale finale del processo che, nel caso presente, risulta costituito da un sistema di diffusione basato sulla tecnologia *DaWinciMD*. Conseguentemente, anche la descrizione di

dettaglio appena fornita per il *workflow* 1 può essere immediatamente riutilizzata per il *workflow* 2 con l'unica avvertenza di sostituire il termine *DaWinciPD* con *DaWinciMD*.
Ciò che invece è opportuno mettere in rilievo è il fatto che le specifiche operazioni effettuate dai motori di ETL e l'architettura della data warehouse intermedia risulteranno diverse nei due casi, in virtù della differente struttura dei sistemi target e delle specifiche organizzazioni dei relativi contenuti informativi statistici. Tutto ciò, come già messo in luce in precedenza, risulta del tutto trasparente rispetto all'utente costruttore che, nella sostanza, distingue i due scenari soltanto per il diverso sistema target selezionato.

3.3. Workflow 3: dai microdati al sistema OLAP-SMOL

In fig.3.2 è rappresentato lo schema di processo del *workflow* 3, relativo al passaggio dai microdati validati (ed un insieme di metadati a corredo) ad un target costituito da un sistema di analisi per attività di tipo OLAP. In questo caso, come sarà chiaro fra poco, non sarà necessario un secondo passo di ETL, dal momento che la data warehouse generata dal primo motore di ETL va di fatto a costituire la componente dati del sistema target. Si riportano di seguito i dettagli del processo:

0. *Attivazione del workflow*. L'utente costruttore intende realizzare un nuovo sistema di analisi di tipo OLAP-SMOL. A tale scopo, entra in contatto con lo specialista Istar, il quale autorizzerà l'utente costruttore all'ingresso nel sistema informatico Istar, secondo un opportuno profilo di accesso.
1. *Generazione e configurazione del sistema target*. L'utente interno installa sulla propria postazione di lavoro l'applicazione SAS OLAP.SMOL. Lo specialista Istar predispone le necessarie infrastrutture hardware e software che ospiteranno il sistema target, dopodiché provvede alla sua installazione e configurazione. Al termine di tali operazioni, l'utente costruttore ha a disposizione un database server che costituisce la componente di back-end del sistema target ed un'applicazione SAS sulla propria postazione di lavoro che ne rappresenta la componente client. Parallelamente a tali attività di competenza dello specialista Istar, l'utente costruttore può già accedere al sistema Istar per compiere le attività di cui ai successivi passi 2 e 3.
2. *Estrazione delle informazioni dal sistema sorgente*. Tramite il motore ETL1 di Istar, l'utente costruttore attiva un processo di estrazione automatica di tutti i metadati statistici disponibili nell'ambiente source, metadati che vengono memorizzati nel database del sistema Istar.
3. *Definizione del dominio statistico*. L'utente costruttore utilizza Istar per visualizzare i metadati estratti dall'ambiente source nel corso del passo 2 e completare eventualmente il proprio dominio statistico, definendo nuove variabili, classificazioni, oggetti di analisi, etc. A fronte dei metadati costituenti il dominio statistico, il sistema Istar attua una fase di mapping automatico che determina l'insieme dei metadati informatici che saranno memorizzati in *database metadati* ed utilizzati nel successivo passo 4 per la generazione ed il caricamento della data warehouse.
4. *Generazione e popolamento della data warehouse*. FOXTROT.MAIN attiva il processo ETL1.DW che, in funzione dei metadati statistici ed informatici contenuti nel database dei metadati, provvede a generare la data warehouse e a popolarla con i dati estratti dal sistema sorgente ed elaborati secondo opportune regole di trasformazione. La struttura risultante è di tipo stellare e costituisce, nella sostanza, la componente dati del sistema OLAP.SMOL. Una volta completate le attività *generazione target* e *definizione dominio*, l'utente costruttore può attivare nuovamente il motore ETL1 di Istar che, in funzione dei metadati statistici ed informatici contenuti nel database dei metadati, provvede a generare

la data warehouse e a popolarla con i dati estratti dal sistema source ed elaborati secondo opportune regole di trasformazione. Tale data warehouse costituisce, di fatto, la componente dati del sistema target OLAP.SMOL che, a questo punto, risulta già disponibile per gli utenti (interni) analizzatori. Ciò significa, in altri termini, che la presente fase consente di ottenere ciò che nei *workflow* di tipo 1 e 2 veniva conseguito attraverso i passi 4 e 5.

5. *Esercizio del sistema target*. L'utente interno accede al sistema target così realizzato ed effettua attività di tipo OLAP sulla data warehouse disponibile.

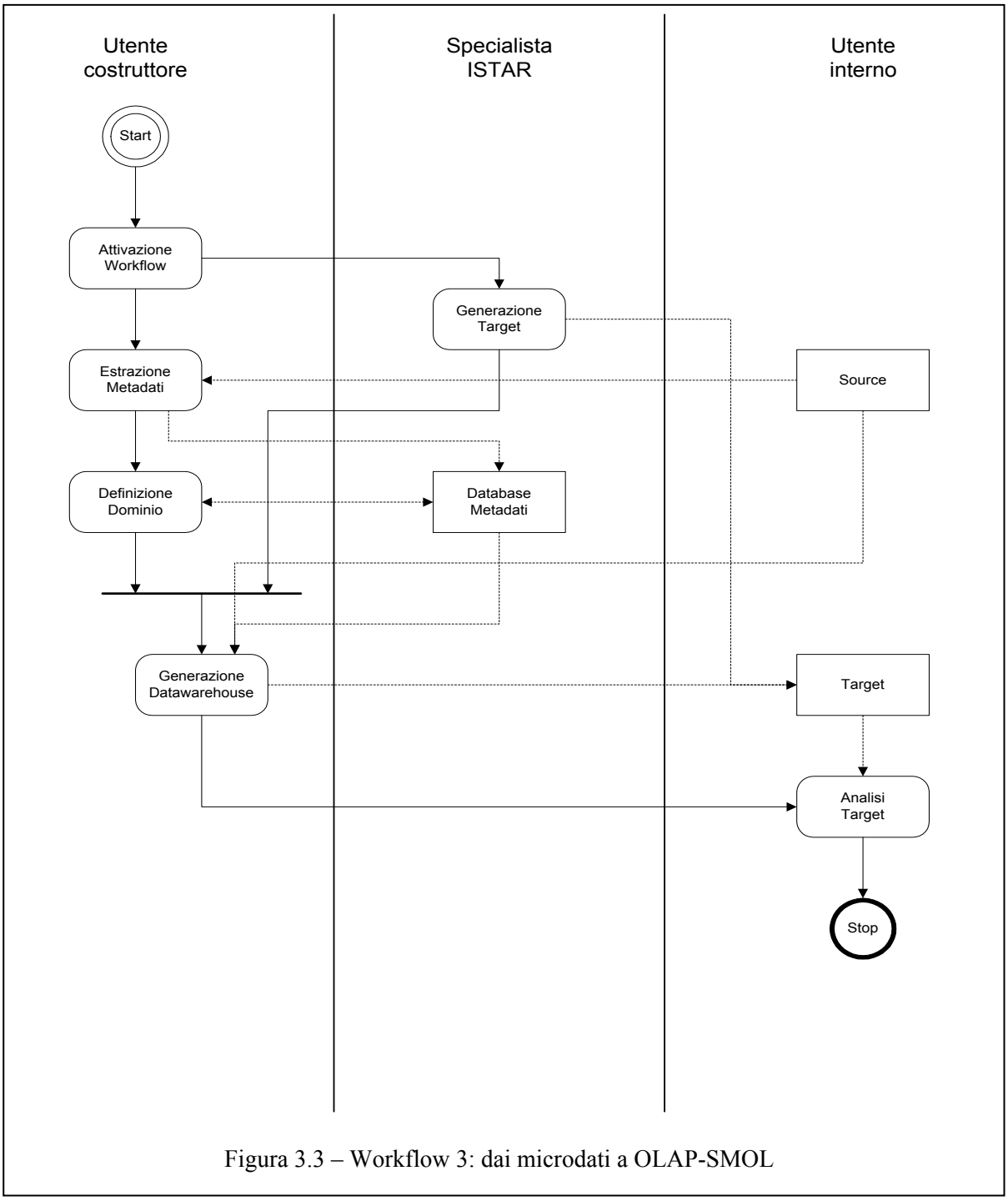

Figura 3.3 – Workflow 3: dai microdati a OLAP-SMOL

3.4. Workflow 4: dalle tavole al sistema DaWinciPD

La quarta ed ultima tipologia di *workflow*, graficamente illustrata in fig.3.3 prende in considerazione lo scenario in cui il sistema sorgente risulta costituito da un semplice insieme di tavole statistiche da diffondere ed il target è un sistema di tipo *DaWinciPD*. In tal caso, il processo risulta semplificato rispetto a tutti quelli precedentemente descritti, sostanzialmente a causa del fatto che tale ciclo di servizio ha origine a valle di tutti i processi di elaborazione che hanno consentito di trasformare i microdati validati nel set di tavole da diffondere e che, nelle tre tipologie di *workflow* precedentemente descritte facevano parte integrante del ciclo di servizio. Si riportano di seguito i dettagli del *workflow*:

0. *Attivazione del workflow*. L'utente costruttore intende realizzare un nuovo sistema di diffusione di tipo *DaWinciPD*. A tale scopo, entra in contatto con lo specialista Istar, il quale autorizzerà l'utente costruttore all'ingresso nel sistema informatico Istar, secondo un opportuno profilo di accesso.
1. *Generazione e configurazione del sistema target*. Lo specialista Istar predispone le necessarie infrastrutture hardware e software che ospiteranno il sistema target, dopodiché provvede alla sua installazione e configurazione. Al termine di tali operazioni, l'utente costruttore ha a disposizione due ambienti fisici distinti: un sistema di diffusione *DaWinciPD* in ambiente intranet (accessibile tramite meccanismo basato su username/password, da utilizzare per il popolamento e la verifica dei contenuti informativi da diffondere) ed uno esposto su Internet (accessibile a tutti e sul quale saranno di volta in volta caricati esclusivamente i contenuti informativi pubblicati). Entrambi i sistemi di diffusione forniti all'utente costruttore sono inizialmente privi di contenuti informativi statistici. Parallelamente a tali attività di competenza dello specialista Istar, l'utente costruttore può già accedere al sistema Istar per compiere le attività di cui ai successivi passi 2 e 3.
2. *Estrazione delle informazioni dal sistema sorgente*. Tramite il motore ETL1 di Istar, l'utente costruttore attiva un processo di conversione automatica di tutte le tavole statistiche da diffondere dal loro formato *proprietario*[16] al formato *standard*[17] del sistema Istar. A questo punto, il sistema target è di fatto pronto per essere popolato e non é prevista, in questo caso, data la natura dell'ambiente source, alcuna fase di definizione di un dominio statistico (implicitamente determinato dal contenuto delle stesse tavole) e di creazione di una struttura di data warehouse intermedia.
3. *Popolamento del sistema target*. L'utente costruttore, tramite il motore ETL2 ed un'opportuna interfaccia utente del sistema Istar, inserisce nel proprio sistema target le tavole statistiche disponibili nel formato XML *standard* e la relativa documentazione a corredo, arricchendone così il contenuto informativo statistico. Tali informazioni vengono via via verificate e validate dallo stesso utente costruttore (tramite il sistema *DaWinciPD* disponibile in ambiente intranet) e rese disponibili alla comunità Internet (utilizzando opportune funzionalità di pubblicazione rese disponibili dal sistema Istar).
4. *Esercizio del sistema target*. L'utente esterno accede via Internet al sistema target *DaWinciPD* così realizzato e fruisce dei contenuti informativi statistici in esso presenti.

([16]) Si tratta, tipicamente, di un formato XML il cui schema deriva direttamente da una operazione di export da ambienti come MS Excel o MS Access.
([17]) Si tratta di documenti XML conformi ad uno schema XSD univocamente definito in ambito Istar.

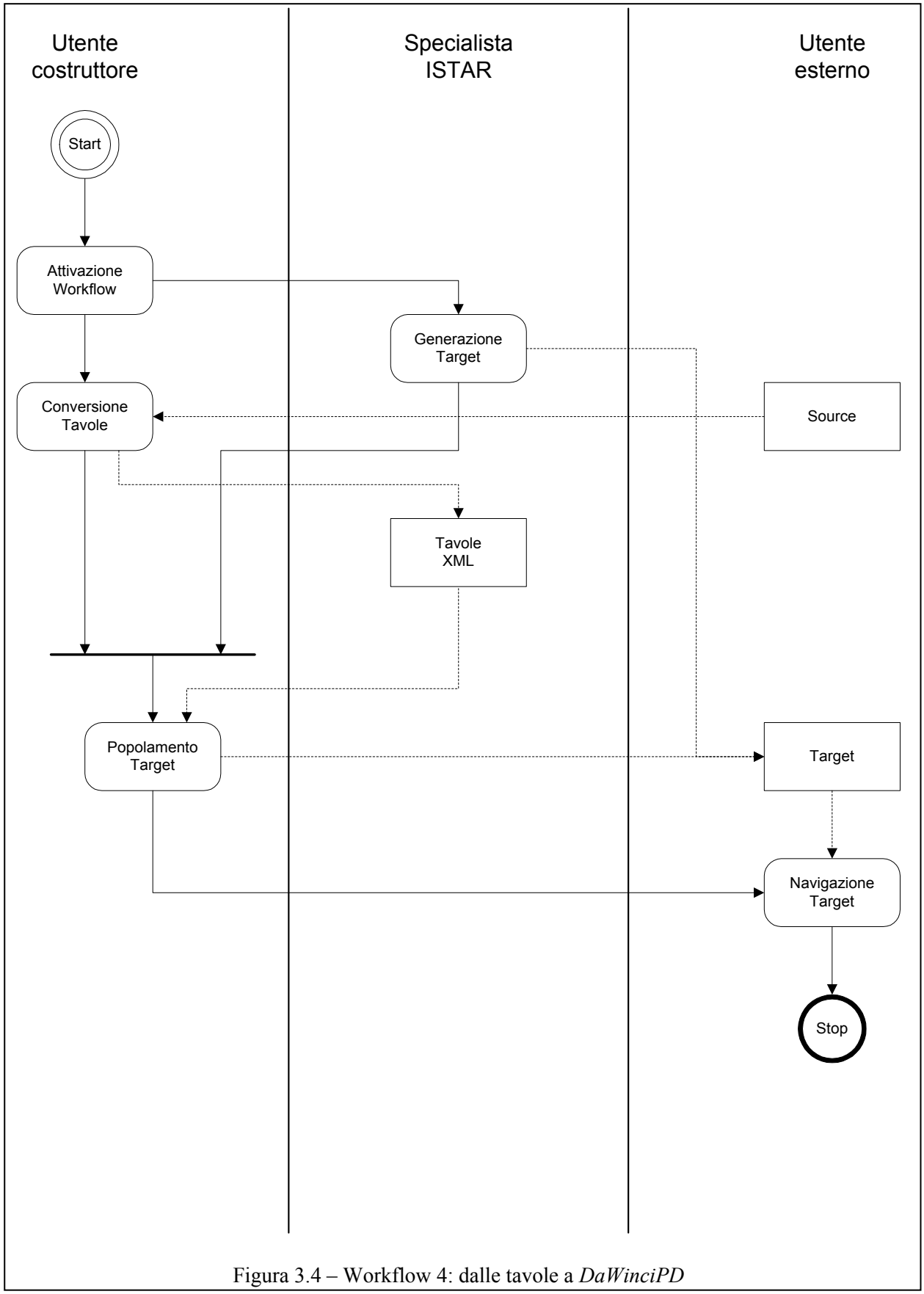

Figura 3.4 – Workflow 4: dalle tavole a *DaWinciPD*

Oltre alle quattro tipologie di *workflow* appena descritte, è prevista la realizzazione di una componente OLAP in ambiente *Business Objects*. In tal caso però si potrebbe dire che il *workflow* – analogo a quello denominato *workflow* 3 e definito nello strato delle funzionalità OLAP in fig.1 come *OLAP micro completa* - si fermerà al momento della generazione dello *strato semantico*. Da quel momento in poi, sarà l'ambiente *Business Objects* stesso a fornire tutti gli strumenti generalizzati di disegno, amministrazione e interrogazione di cui dispone[18].

4. Architettura applicativa e dispiegamento tecnologico di Istar

Nella presente sezione saranno illustrate le caratteristiche generali dell'architettura del sistema Istar. In primo luogo, saranno descritte la struttura del software e della componente dati del sistema. Successivamente, tramite una descrizione dei flussi elaborativi interni al sistema, saranno analizzate le relazioni fra le possibili tipologie di *workflow* descritte nella precedente sezione §.3 e le componenti software del sistema Istar. Infine, saranno forniti dettagli sul dispiegamento fisico del sistema, ovvero sull'architettura hardware e software utilizzata come piattaforma di esercizio.

4.1. Struttura del software

Il software Istar presenta una struttura costituita da quattro componenti principali, che vengono di seguito brevemente illustrate.

INTERFACCIA DI GESTIONE

Si tratta del modulo, denominato *Foxtrot*, che costituisce l'interfaccia utente verso il sistema e che contiene la logica applicativa che consente di:

- ✓ rappresentare il dominio statistico di interesse,
- ✓ trasformare i concetti propri del dominio statistico nei corrispondenti concetti propri della letteratura sui sistemi di data warehouse,
- ✓ gestire tutte le attività del *workflow* che portano alla realizzazione del sistema target.

Tale componente presenta un'architettura sostanzialmente di tipo three-tiers[19].

MOTORE DI ETL

Motore di estrazione, trasformazione e caricamento che consente, nel suo complesso, il passaggio dal dato nel suo formato sorgente (microdato validato o tavola statistica) al dato da diffondere o su cui effettuare attività di tipo OLAP.

([18]) Obiettivo di più lungo periodo rispetto al progetto in corso, è quello di personalizzare l'ambiente Business Objects per rendere agevole soprattutto i meccanismi di *manutenzione semantica,* cioè la gestione dinamica delle dimensioni di analisi, delle gerarchie di navigazione, nonché della stessa struttura multidimensionale di riferimento, sulla falsariga di quanto già realizzato in alcune esperienze ISTAT, come ad esempio la Data Warehouse sulle Cause di Morte, già operativa da anni.

([19]) Ad eccezione del modulo di gestione del ciclo di vita delle tavole statistiche nell'ambito del *workflow* 2, che è stato realizzato secondo il paradigma client-server.

Tale componente consiste, tipicamente, in un insieme di moduli software direttamente residenti sul database server.

SISTEMI DI ANALISI

Sistemi di diffusione in ambiente intranet e riservati, quindi, ad un'utenza interna, orientati ad attività di tipo OLAP sia sul microdato che sul dato aggregato. Allo stato attuale, l'unico sistema disponibile, OLAP.SMOL, è stato realizzato in ambiente SAS ed in architettura client-server.

SISTEMI DI DIFFUSIONE

Sistemi di diffusione destinati all'utenza esterna e basati sulla tecnologia DaWinci, nei suoi due principali filoni attualmente disponibili: *DaWinciPD* (per la diffusione di un insieme predefinito di tavole statistiche navigabili rispetto al tempo, al territorio e ad un albero gerarchico di categorie) e *DaWinciMD* (per la diffusione di dati statistici navigabili secondo la filosofia oggetto-classificazione).
In entrambi i casi, il paradigma architetturale adottato è quello *three-tiers*.

Informazioni di maggior dettaglio sull'architettura del software Istar sono riportate in appendice al presente documento.

4.2. Architettura dei dati

In linea con quanto finora esposto, la componente dati del sistema Istar può essere considerata partizionata nelle quattro basi di dati di seguito descritte:

- ✓ *Database dei metadati*: contiene tutti i metadati (statistici, informatici e di configurazione) necessari al sistema per la corretta e completa esecuzione del *workflow* e, conseguentemente, per la predisposizione del sistema target.
- ✓ *Data warehouse*: struttura dati intermedia fra ambiente source e sistema target nei casi *DaWinciPD* e *DaWinciMD*, ovvero componente dati del sistema target nel caso OLAP-SMOL.
- ✓ *Web warehouse interna del sistema target*: componente dati del sistema target disponibile in ambiente intranet; include tutto il contenuto informativo prodotto (tavole statistiche e relativa documentazione) ed i metadati necessari per l'accesso a tale contenuto informativo.
- ✓ *Web warehouse esterna del sistema target*: componente dati del sistema target esposta su Internet; consiste in una replica parziale della web warehouse interna ed include tutto il contenuto informativo pubblicato (tavole statistiche e relativa documentazione) ed i metadati necessari per l'accesso a tale contenuto informativo[20].
  delle strutture di produzione

4.3. Flussi elaborativi

A questo punto, nota la struttura generale del software e della componente dati del sistema, risulta possibile analizzare i *workflow* precedentemente descritti concentrando l'attenzione

([20]) Ovviamente, i due web warehouse appena descritti esistono soltanto nei casi *DaWinciPD* e *DaWinciMD* e non nel caso OLAP-SMOL.

sul flusso dell'informazione statistica all'interno del sistema e sul ruolo che le varie componenti software rivestono all'interno dei vari cicli di servizio.
In fig.4.1, ad esempio, è rappresentato il flusso elaborativo associato al *workflow* di tipo 1. In questo caso, il flusso dell'informazione statistica ha origine dall'ambiente source, sistema dal quale, tramite il modulo ETL1, viene in un primo momento estratta tutta la metainformazione di interesse, la quale viene memorizzata nel database dei metadati. Grazie a tali metadati ed ai microdati estratti ancora dal sistema sorgente, è lo stesso modulo ETL1 che si occupa poi di generare e popolare la data warehouse grazie alla quale sarà possibile alimentare il sistema target. A questo punto, è il modulo ETL2 che, proprio a partire dalla data warehouse, si occupa di generare le tavole statistiche che andranno a far parte del sistema target di diffusione, rendendo così l'informazione statistica fruibile da parte dell'utente esterno. *Foxtrot*, in tutto questo, è il modulo integrato attraverso il quale l'utente costruttore può interagire con il sistema Istar, sia per attivare i due motori di ETL su cui fa perno il flusso elaborativo, sia per la gestione della metainformazione necessaria per il compimento del ciclo di servizio (metadati statistici per la definizione del dominio di analisi e metadati di configurazione per l'organizzazione dei contenuti informativi statistici del sistema target).

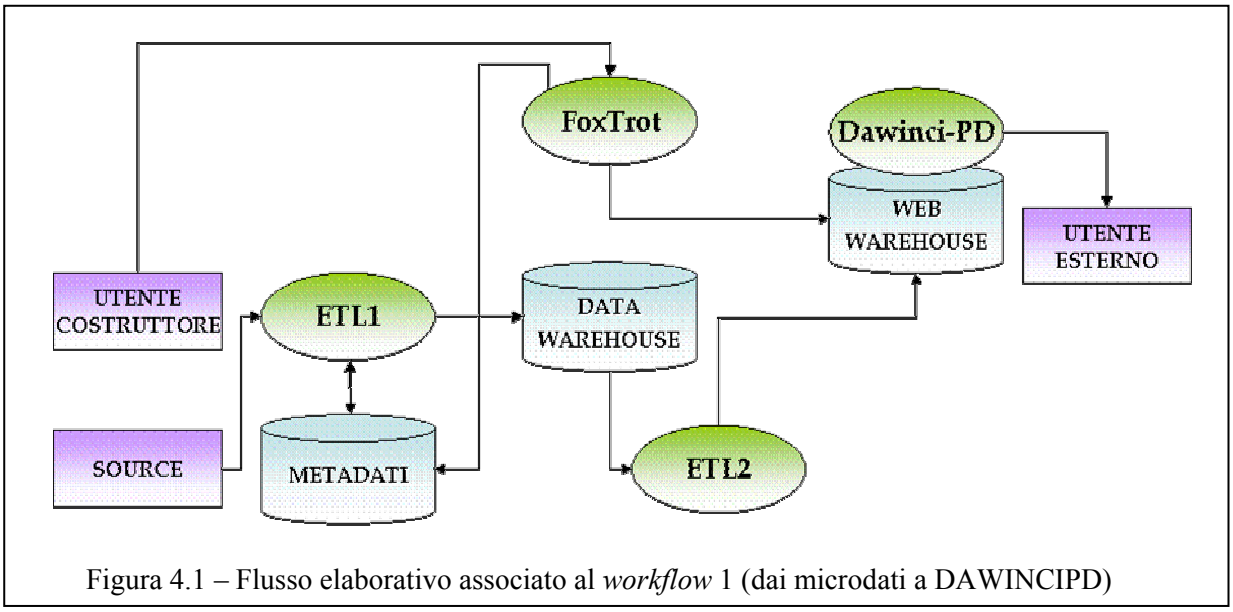

Figura 4.1 – Flusso elaborativo associato al *workflow* 1 (dai microdati a DAWINCIPD)

In modo del tutto analogo può essere descritto il flusso elaborativo associato al *workflow* 2 (cfr. fig.4.2). L'unica differenza, come del resto risulta chiaro anche dalla descrizione dei rispettivi *workflow*, consiste nel fatto che in questo caso il sistema target si basa sulla tecnologia *DaWinciMD*, il che, come già messo in evidenza, comporta l'esecuzione di specifiche operazioni. In particolare, il comportamento del modulo ETL1, unico nei due casi, differisce nella componente di generazione dello schema fisico della data warehouse che necessita di strutture dati diverse in funzione della diversità dei sistemi target. Per quanto riguarda il secondo passo di ETL2, vengono nei due casi attivati due moduli software distinti e realizzati ad hoc in funzione dei due possibili sistemi di diffusione. Ovviamente, tali differenze elaborative vengono automaticamente determinate dal sistema Istar e risultano del tutto trasparenti rispetto agli utenti del sistema.

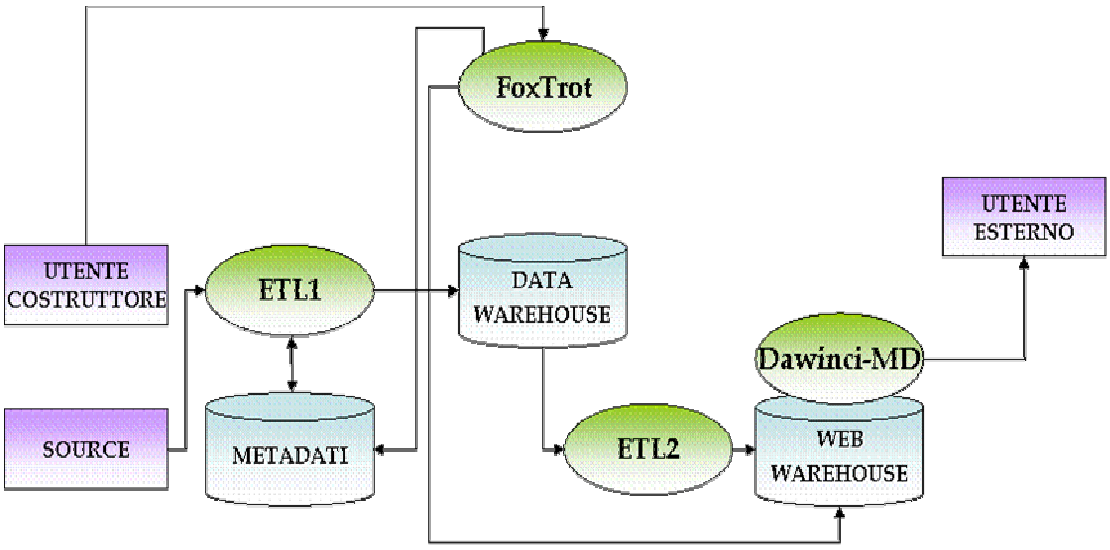

Figura 4.2 – Flusso elaborativo associato al *workflow* 2 (dai microdati a *DaWinciMD*)

Come appare evidente dalla fig.4.3, il flusso elaborativo associato al *workflow* di tipo 3 risulta semplificato rispetto a quelli sin qui analizzati. In particolare, il flusso è caratterizzato da un solo passo di ETL, operato dal modulo ETL1, con caratteristiche del tutto analoghe a quelle relative ai *workflow* di tipo 1 e 2 e con la sola specificità dovuta alla componente di generazione dello schema fisico della data warehouse, la cui struttura risulta in generale diversa da quelle necessarie per i sistemi di diffusione. Non esiste, in questo caso, un secondo passo di ETL, dal momento che la data warehouse stessa costituisce, come già evidenziato in precedenza, la componente dati del sistema target OLAP-SMOL, componente sulla quale l'utente interno potrà direttamente effettuare le sue attività di tipo OLAP.
Nel caso in esame, il modulo *Foxtrot* viene utilizzato sostanzialmente come interfaccia per il completamento della definizione del dominio di analisi.

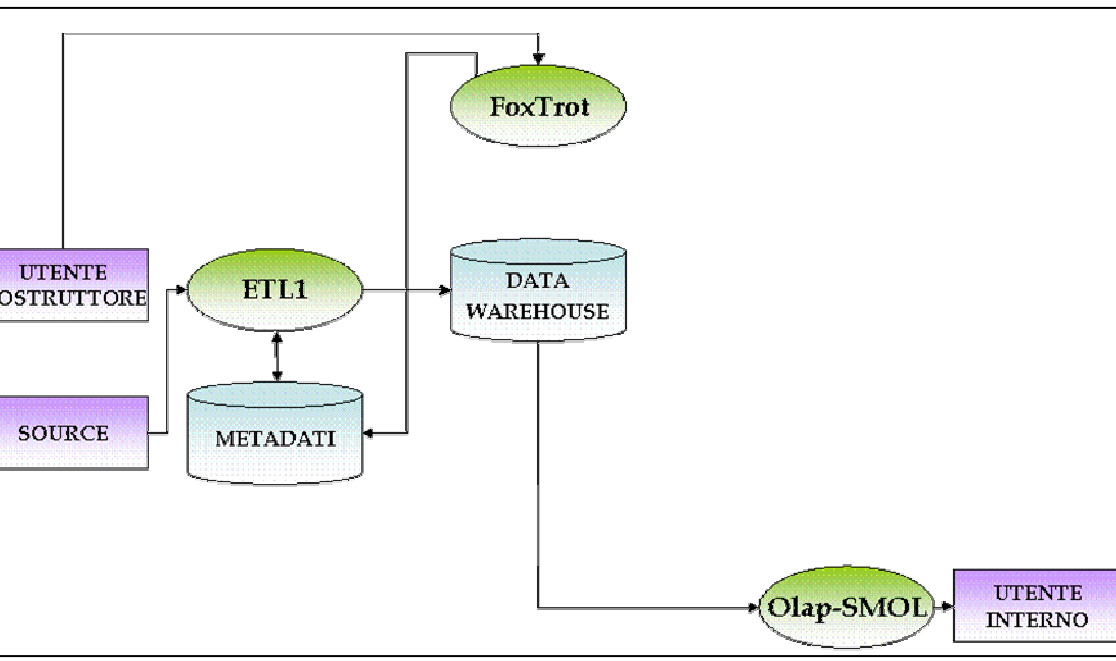

Figura 4.3 – Flusso elaborativo associato al *workflow* 3 (dai microdati a OLAP-SMOL)

L'ultima tipologia di flusso elaborativo, associata al *workflow* 4, che consente di realizzare un sistema target di tipo *DaWinciPD* a partire da un insieme di tavole statistiche, è rappresentata graficamente in fig.4.4. In uno scenario di questo tipo, il processo risulta ulteriormente semplificato e non richiede la generazione di una vera e propria struttura intermedia di data warehouse, ma, come già sottolineato, una più banale conversione delle tavole da diffondere, dal loro formato sorgente *proprietario* nel formato *standard* definito in ambito del sistema Istar. A questo punto, il modulo ETL2 consiste in una semplice procedura che si occupa di caricare le singole tavole, disponibili in formato XML *standard*, nel sistema target di diffusione.

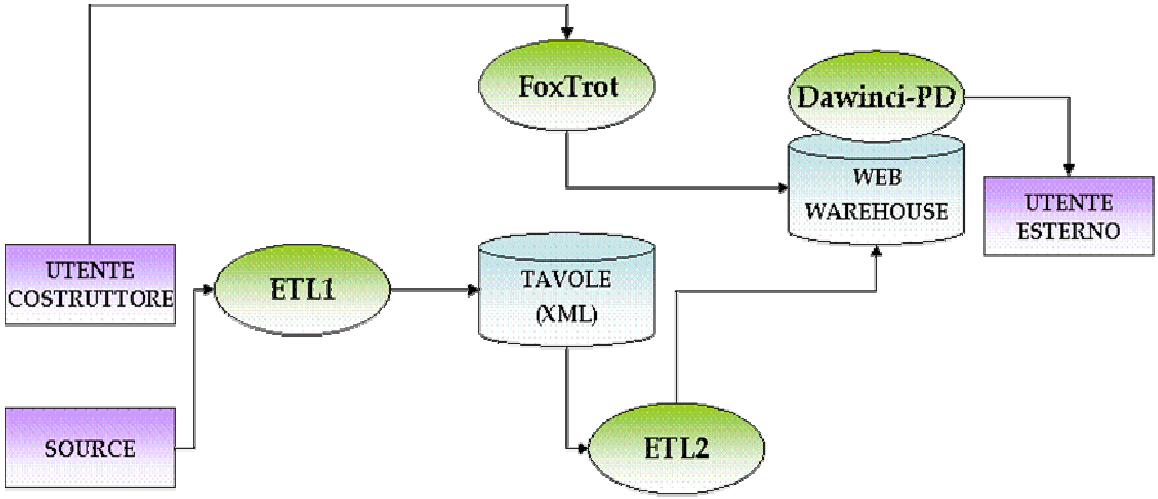

Figura 4.4 – Flusso elaborativo associato al *workflow* 4 (dalle tavole a *DaWinciPD*)

In quest'ultimo caso, infine, data la natura dell'ambiente source considerato, non risulta necessaria una vera e propria gestione dei metadati statistici; pertanto, il modulo *Foxtrot* viene utilizzato sostanzialmente per la gestione dei metadati di configurazione del sistema target.

4.4. Dispiegamento fisico

Analogamente a quanto già descritto per la struttura dei *workflow* e per i flussi di elaborazione, anche per quanto riguarda il dispiegamento fisico, è possibile individuare quattro differenti tipologie di architettura che, come si vedrà fra poco, vengono sostanzialmente determinate dal tipo di sistema sorgente e target coinvolti.
Prendendo in considerazione il *workflow* di tipo 1 (dai microdati a *DaWinciPD*), il relativo dispiegamento fisico (cfr. fig.5.1) prevede la presenza di tre host (aventi sia funzioni di database server che di application server) su cui risiedono, rispettivamente, il sistema Istar ed i due sistemi di diffusione di tipo *DaWinciPD* (quello disponibile in ambiente intranet, per la produzione e verifica dei contenuti informativi statistici, e quello esposto su Internet, contenente le sole informazioni già validate per la pubblicazione). Per quanto concerne il modulo *Foxtrot* di interfaccia utente con il sistema, è possibile distinguere tre componenti distinte: la prima residente sul server Istar e relativa ai metadati statistici (relativi alla definizione del dominio statistico), la seconda sul server di diffusione intranet per la gestione dei metadati (di configurazione e di documentazione) del sistema target, la terza ed ultima,

installata sul client dell'utente costruttore, dedicata alla gestione del ciclo di vita (definizione, produzione e pubblicazione) delle tavole statistiche.
Sia l'utente costruttore sia l'utente interno, infine, hanno a disposizione un semplice browser tramite il quale effettuare, rispettivamente, la verifica dei contenuti informativi statistici prodotti (sul sistema di diffusione intranet) e la navigazione di quelli già pubblicati (sul sistema di diffusione Internet).

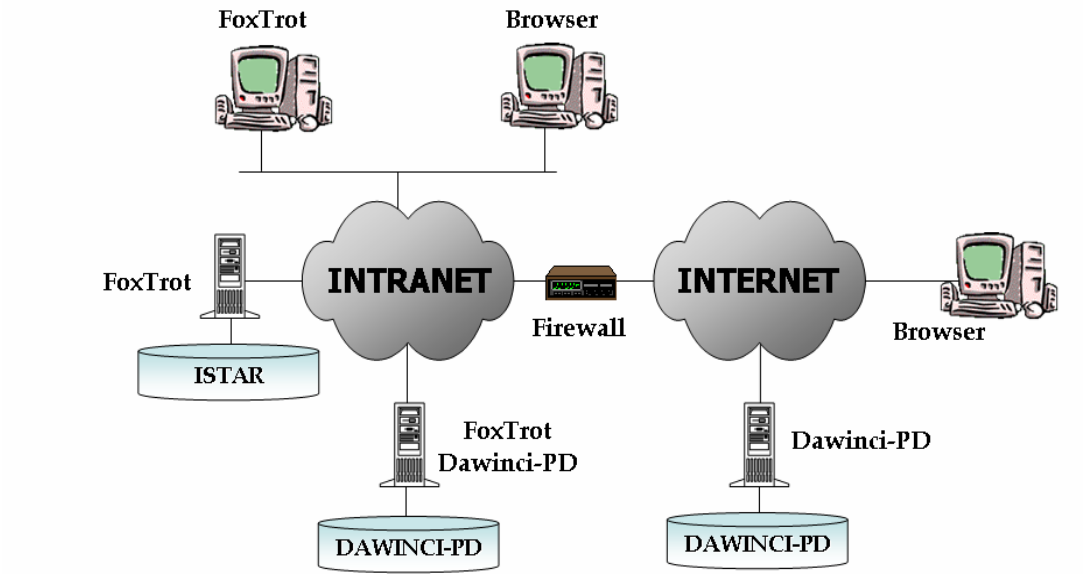

Figura 5.1 – Dispiegamento fisico relativo al *workflow* 1 (dai microdati a *DaWinciPD*)

Il dispiegamento fisico relativo al *workflow* 2 (dai microdati a *DaWinciMD*) non risulta molto diverso da quello appena descritto (cfr. fig.5.2). Sono presenti anche in questo caso, infatti, i tre host dedicati al sistema Istar ed ai due sistemi di diffusione (che, in questo caso, sono evidentemente basati sulla tecnologia *DaWinciMD*). Risulta invece diverso il dispiegamento del modulo *Foxtrot*, il quale prevede la presenza di due sole componenti residenti, rispettivamente, sul server Istar e sul server di diffusione intranet. Ciò si deve banalmente al fatto che, nel caso *DaWinciMD*, tutte le funzionalità di gestione dei metadati presentano un'architettura di tipo *three-tiers* e risiedono, pertanto, sulla componente application server del sistema di diffusione.

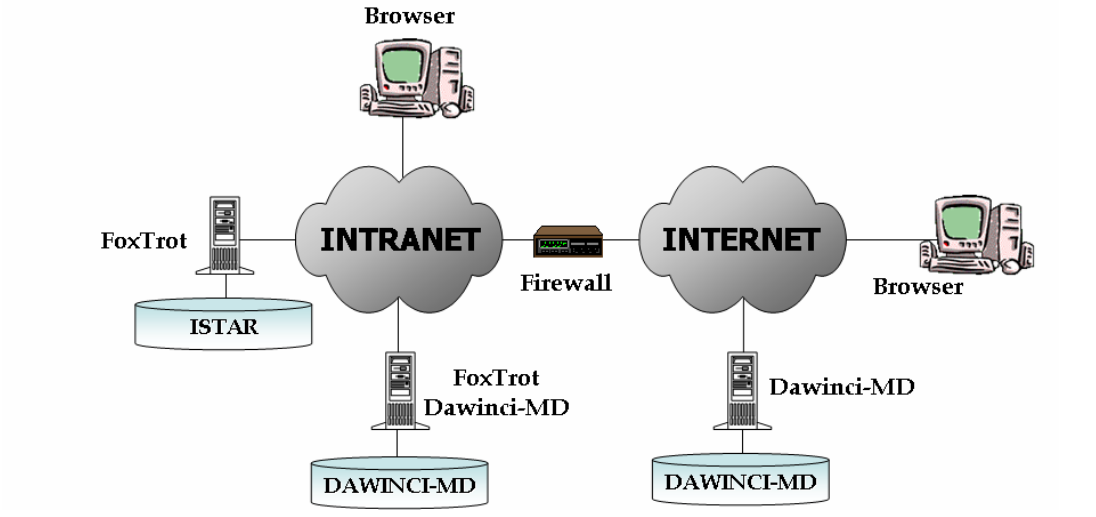

Figura 5.2 – Dispiegamento fisico relativo al *workflow* 2 (dai microdati a *DaWinciMD*)

La realizzazione del *workflow* di tipo 3 (dai microdati a OLAP-SMOL) prevede un dispiegamento fisico sensibilmente diverso dai due sin qui analizzati (cfr. fig.5.3). In questo caso, innanzitutto, non è prevista la presenza di alcuna componente esposta su Internet, dal momento che l'utente finale del sistema target può essere esclusivamente un utente interno all'Istituto. Inoltre, l'architettura software di OLAP-SMOL, contrariamente a quanto visto a proposito dei sistemi di diffusione, é di tipo client-server e prevede, pertanto, la sola presenza di un database server e di una postazione client per l'utente interno. Resta confermata, infine, la presenza del server Istar e della componente *Foxtrot* ivi residente per la gestione dei metadati di interesse per il caso specifico.

Per quanto riguarda il *workflow* di tipo 4 (dalle tavole a *DaWinciPD*), infine, si può fare riferimento a quanto già detto a proposito del *workflow* 1, con un paio di eccezioni determinate dal fatto che, nel caso qui in esame, il punto di partenza risulta essere un dato già elaborato ed aggregato in un set di tavole pronte per la diffusione e non un insieme di microdati validati ancora da lavorare. Ciò determina, nella sostanza, l'assenza di un vero e proprio database server Istar (dal momento che risulta in questo caso priva di senso la gestione di un dominio statistico) e della componente *Foxtrot* client-server che, nell'ambito del *workflow* di tipo 1, consentiva la gestione del ciclo di vita delle tavole statistiche. Nello scenario qui in esame, tale componente viene più banalmente sostituita dal client tramite il quale viene prodotto il set di tavole da diffondere (a titolo di esempio, in fig.5.4 si fa esplicito riferimento al prodotto MS Excel).

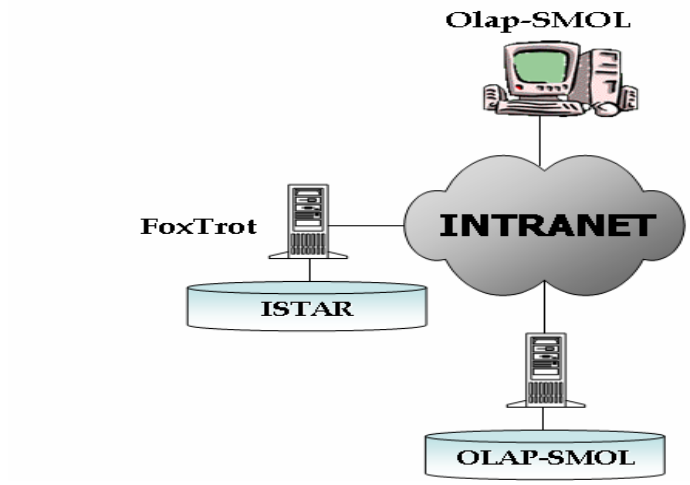

Figura 5.3 – Dispiegamento fisico relativo al *workflow* 3 (dai microdati a OLAP-SMOL)

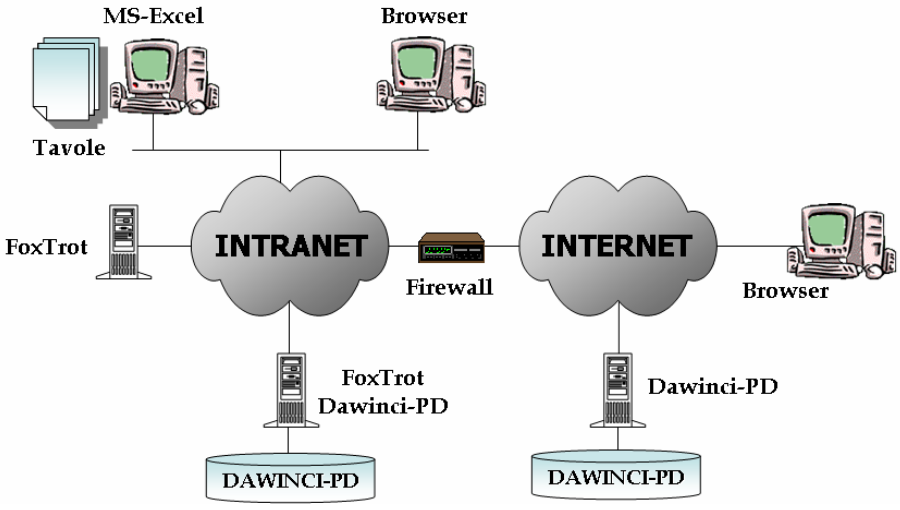

Figura 5.4 – Dispiegamento fisico relativo al *workflow* 4 (dalle tavole a *DaWinciPD*)

5. Sviluppi futuri

Per quanto riguarda lo sviluppo delle componenti aggiuntive interne ad Istar si individuano nei seguenti punti le problematiche più importanti per l'evoluzione del sistema:

- gestione tavole multioggetto
- gestione serie storiche e gerarchie temporali
- gestione classificazioni tematiche *pivot* (ad esempio Ateco per strutturali, cause di morte, spese per consumi, ecc. ecc.)
- realizzazione nuovo data warehouse per aggregazioni on-line o, comunque, indipendenti da tavole predefinite
- generazione di modalità di accesso al sistema di navigazione con *filtri territoriali*

Tra le attività indicate, alcune sono in avanzata fase di sviluppo. Tra queste riveste particolare importanza la gestione delle tavole multioggetto. In sostanza, con la realizzazione di un meccanismo di "gestione di tavole multioggetto", si fa riferimento ad un ampliamento delle funzionalità di navigazione che consenta di visualizzare nella stessa pagina vari "oggetti" (ad esempio in ambito Censimento la popolazione residente e la popolazione presente oppure indicatori come il numero di anziani per un bambino con l'indice di vecchiaia e la percentuale di popolazione con più di xx anni). In tal modo si offre all'utente la possibilità di costruire, in modo flessibile e non precostituito, indicatori di suo interesse a partire dai dati assoluti riferiti agli oggetti selezionati. L'applicazione è sviluppata in modo tale da permettere un ulteriore grado di libertà: è possibile caratterizzare uno degli oggetti selezionati in base ad una classificazione ad esso riferita. In tal modo, l'utente costruirà una pagina che sarà composta affiancando i dati riferiti ad entrambi gli oggetti, di cui uno potrà essere dettagliato in base alla classificazione scelta.
Altra funzionalità chiave sarà costituita dal sistema di navigazione con filtri territoriali. L'idea alla base è che sia consentita una selezione del territorio preventiva (eventualmente con possibilità di *login* e *password*), dopodiché l'utente potrà navigare secondo le regole abituali del sistema, ma solo relativamente al territorio prescelto. Questa versione del sistema è anche legata alla produzione dei CD "personalizzati" da allegare a volumi territoriali, in quanto la materializzazione del sito si deve basare su tale versione del sistema, in modo che l'utente non possa muoversi sulla gerarchia territoriale una volta giunto sulla tavola statistica

6. Conclusioni

Nel presente lavoro abbiamo illustrato come nella costruzione del Sistema informativo generalizzato di diffusione dell'Istat si sia tenuto conto congiuntamente di due aspetti considerati solitamente distinti all'atto della progettazione dei S.I., rappresentati dalle funzioni applicative e dai processi aziendali. Si è cercato cioè di porre l'attenzione non solo al miglioramento della qualità dei prodotti e dei servizi *direttamente visibili* dagli utenti finali, ma anche a quello dei *processi* che a tali prodotti e servizi conducono.

È stato anche illustrato come il sistema, giungendo a valle di una serie di esperienze rivolte allo sviluppo di sistemi informativi di diffusione più specializzati e mirati ad alcune specifiche problematiche, quali l'analisi interattiva su microdati, la diffusione di dati statistici sotto forma di tavole predefinite, la costruzione di una originale ontologia del dato aggregato

per la sua diffusione su Web, ecc., si proponga come un sistema di diffusione ad ampio spettro, avente la finalità, da una parte, di accogliere tutte le *nuove* esigenze di diffusione provenienti dai settori di produzione statistica e, dall'altra, di integrarsi con le altre numerose esperienze di sistemi di diffusione esistenti in Istituto, nessuna delle quali andrà persa, e che anzi potrà trovare ampio risalto e valorizzazione proprio entrando a far parte di uno scenario. armonizzato. Questa impostazione evolutiva del processo di sviluppo dei sistemi informativi è stata resa possibile dall'uso, introdotto sin dall'inizio, di tecnologie di progettazione e sviluppo orientate agli oggetti a fianco di metodologie di sviluppo di *workflow*, in modo da favorire il riuso dei *pattern* di progettazione e dei moduli software, nel rispetto delle regole aziendali sottostanti che ogni sistema si trova a fronteggiare, velocizzando di conseguenza le attività di sviluppo integrato di nuovi servizi e prodotti di diffusione relativi al patrimonio di informazioni statistiche disponibili presso l'Istituto.

RINGRAZIAMENTI

Il presente lavoro è frutto di attività di analisi, progettazione e sviluppo condotte congiuntamente da risorse afferenti a diverse strutture dell'Istituto e facenti parte del Gruppo di Lavoro incarico dello sviluppo di Istar e del sistema informativo generalizzato di diffusione. In particolare gli autori rivolgono un doveroso ringraziamento a Maria Cristina Bedeschi, Roberta Benedetti, Francesco Bisceglia, Claudia Cianfarani, Paolo Giacomi, Fulvio Giannetti, Paola Giorgetti, Andrea Marsico, Antonio Laureti Palma, Franco Lodovici, Paolo Piergentili per il prezioso contributo fornito in seno a tale progetto.

BIBLIOGRAFIA

[1] Shoshani, A. (1997) – *OLAP and statistical databases: similarities and differences* – 16th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, 1997

[2] Sindoni G., Tininini L., *A Statistical Web Warehouse System*, Proceedings of Q2004 - European Conference on Quality and Methodology in Official Statistics 2004

[3] Cabibbo L., Torlone R., *The Design and Development of a Logical System for OLAP*, Proceedings of DaWaK 2000 - Int. Conf. on Data Warehousing and Knowledge Discovery, 2000

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

[5] Casati F, Pernici B., *Linguaggi per la modellazione dei processi aziendali*, marzo 1999

[6] De Petra G., *La fase alta della progettazione dei sistemi informativi: dall'analisi integrata di organizzazione e tecnologia alla individuazione e valutazione di costi e benefici della soluzione innovativa*, in Ercoli, P., Batini, C., Marozza, F., *Nuove metodologie per i sistemi infomativi della pubblica amministrazione*, eds., Il Mulino 1994

[7] Hollingsworth D., *Management Coalition The Workflow Reference Model*, Document Number TC00-1003, WFMC, 1995

[8] Flores F., Medina-Mora R., Winograd T., Flores R., "*The action Workflow Approach to Workflow Management Tecnology*", Proc. of the ACM Conference on Computer Supported Cooperative Work, Nov, 1992

[9] Toma A., Bergamasco S., *Action Workflow Analysis - Conversation For Action e i nuovi processi per la gestione della rete di rilevazione per l'indagine sulle Forze di Lavoro*, Quinta Conferenza Nazionale di Statistica, Roma, 15- 17 novembre 2000

[10] Chappell D.A., Jewell T., *Java Web Services*, Edizioni O'Reilly, 2002

[11] Violi P., *Significato ed Esperienza*, Studi Bompiani, Milano, 1997

[12] Petöfi, J. S. *On the Problem of Co-textual Analysis of Texts*, in *International Conference on Computational Linguistics*, Sånga-Säby, Sett. 1969, preprint