Istituto Nazionale di Statistica

Principi Guida per il Miglioramento della Qualità dei Dati Toponomastici nella Pubblica Amministrazione

L’intero volume è stato redatto e coordinato da Piero Demetrio Falorsi, e Monica Scannapieco, ad eccezione del capitolo 2 che è stato redatto insieme ad Antonia Boggia e del capitolo 3 che è stato redatto con Antonio Pavone.

Ciascuno dei capitoli del presente volume è stato coordinato come segue:

- Il Capitolo 1 è a cura di Alessandro Alessandroni, Piero Demetrio Falorsi e Monica Scannapieco.
- Il Capitolo 2 è a cura di Antonia Boggia, Orietta Gargano e Alessandro Pallara.
- Il Capitolo 3 è a cura di Antonio Pavone con il supporto di Lamberto Pizzicannella.
- Il Capitolo 4 è a cura di Antonia Boggia, Fabio Crescenzi, Orietta Gargano.
- Il Capitolo 5 è a cura di Piero Demetrio Falorsi e Monica Scannapieco.
- Il Capitolo 6 è a cura di Alessandro Alessandroni, Piero Demetrio Falorsi e Monica Scannapieco.

Lamberto Pizzicannella ha curato le appendici: Il sistema SISTER per il riconoscimento dei dati toponomastici e L’estrazione del campione, Monica Scannapieco ha curato l’appendice: I file XLM schema, Orietta Gargano ha curato l’appendice: Dizionari, Antonia Boggia e Orietta Gargano hanno curato l’appendice 1: Glossario (allegata al volume), Antonia Boggia e Gabriele Ciasullo hanno curato l’appendice 2: La normativa (allegata al volume), Marcello D’Orazio, Piero Demetrio Falorsi, Antonio Pavone, Marina Signore, Giorgia Simeoni hanno curato l’appendice 3: Definizione di indicatori di qualità per i dati toponomastici (tale appendice, allegata al volume, costituisce il documento di studio preliminare al capitolo 3).

1 Introduzione

Il presente volume illustra l'insieme dei *principi guida* che le pubbliche amministrazioni italiane possono seguire al fine di *migliorare la qualità delle proprie basi di dati toponomastiche.* Tali principi guida sono stati elaborati a partire dai risultati del progetto sviluppato nell'ambito dell'*accordo di collaborazione per la definizione di criteri guida per la gestione della qualità dei dati nella pubblica amministrazione*, siglato il 23 aprile 2002 dall'Autorità per l'Informatica nella Pubblica Amministrazione (AIPA, ora CNIPA) e dall'Istituto Nazionale di Statistica (ISTAT).

Il progetto nasce dalla esigenza di gestire la problematica della qualità dei dati nella Pubblica Amministrazione. La carenza di una soddisfacente qualità dei dati nelle basi di dati pubbliche riguarda molte tipologie di dati. Negli anni, l'acquisizione e la memorizzazione dei dati da parte delle amministrazioni pubbliche è avvenuta in maniera spesso non controllata, senza un adeguato utilizzo degli strumenti atti a consentire la necessaria gestione e manutenzione. Il risultato di tale processo è stato l'accumulo di dati duplicati, incompleti o poco accurati. Nel progetto, tra le diverse tipologie di dati gestiti dalle amministrazioni pubbliche nell'ambito dei processi amministrativi, ci si è focalizzati sulla categoria dei *dati toponomastici*, ossia l' insieme di componenti che identificano in modo univoco un luogo fisico sul territorio.

Le motivazioni che hanno condotto alla scelta di concentrare il lavoro sui dati toponomastici sono: la centralità di tali dati nel contesto delle basi di dati pubbliche e l'assenza di uno standard di rappresentazione dei dati toponomastici.

I dati toponomastici sono memorizzati in una parte preponderante delle basi di dati delle amministrazioni pubbliche ed hanno, dunque, un ruolo centrale nell'assetto dati nazionale. Si consideri, infatti, il ruolo che i dati toponomastici hanno in relazione alla localizzazione dei soggetti fisici e giuridici: l'informazione toponomastica, nella accezione comune di indirizzo, è molto spesso associata alle informazioni identificative dei soggetti fisici e giuridici. Gli indirizzi sono coinvolti in una pluralità di processi amministrativi, la cui buona riuscita è seriamente ostacolata dalla presenza di indirizzi di scarsa qualità che può impedire l'esatta localizzazione e il raggiungimento dei soggetti coinvolti nei processi stessi.

L'assenza di uno standard di rappresentazione dei dati toponomastici si riflette in maniera fortemente negativa sulla qualità di tali dati. I dati toponomastici, infatti, sono caratterizzati dall'avere una struttura complessa, definita cioè da più componenti distinte. L'individuazione di tali componenti non è univoca e le diverse componenti presentano al loro interno relazioni che possono non essere sempre facilmente definibili. Inoltre, esistono diverse forme possibili di rappresentazione dei dati toponomastici, con una conseguenza immediata sulle caratteristiche di qualità. Ad esempio, un dato toponomastico può essere memorizzato come un unico elemento oppure suddiviso in varie componenti (toponimo, denominazione etc.): la memorizzazione come elemento unico spesso ingenera l'introduzione di errori, dovuti alla mancata specifica di una componente o all'inserimento erroneo della componente stessa.

1.1 I Contenuti e gli Utenti dei Principi Guida

Il presente volume propone un quadro metodologico di riferimento per le amministrazioni che vogliono intraprendere azioni di miglioramento della qualità dei propri dati toponomastici. Tale quadro si fonda sul principio base di garantire e potenziare la cooperazione tra le pubbliche amministrazioni. Nel contempo, esso considera anche le realtà interne alle amministrazioni, nel rispetto della loro autonomia e indipendenza.

I contributi fondanti di tale quadro metodologico sono di seguito elencati:

- la definizione di formati di acquisizione ed interscambio dei dati toponomastici, a partire da una concettualizzazione delle varie componenti del dato toponomastico. Con la proposta di tali formati, si offre una soluzione al problema della rappresentazione del dato toponomastico, consentendo ad ogni amministrazione di effettuare adeguati controlli prima di memorizzare i dati toponomastici, sia nella fase di acquisizione diretta dal cittadino, che nella fase di acquisizione e scambio con un'altra amministrazione;
- la proposta di intervento sui processi di interscambio dell'informazione toponomastica in modo da garantire l'allineamento di tale informazione nelle basi di dati pubbliche. Infatti, l'informazione toponomastica è intrinsecamente variabile nel tempo; ad esempio, sono diffusi gli eventi di ridenominazione di strade, di fusione di comuni, etc. Allo stato attuale, tali eventi non vengono propagati a tutte le basi di dati interessate, causando un disallineamento dell'informazione memorizzata e la presenza di informazione errata in quanto non aggiornata;
- la definizione di una metodologia di misurazione e miglioramento della qualità dei dati toponomastici. Un'attività di misurazione è necessaria alle amministrazioni sia per orientare in maniera opportuna interventi di miglioramento, sia per dichiarare la qualità delle proprie basi di dati alle altre amministrazioni, così favorendo e potenziando gli scambi tra amministrazioni. Ad esempio, nell'ambito della attività di misurazione, sono state individuate delle aree critiche, ottenute come aggregazione di territori comunali, considerati nel loro complesso. Nelle basi dati della Pubblica Amministrazione, i dati toponomastici riferiti a tali aree possono presentare problemi di scarsa qualità. L'identificazione delle aree critiche rappresenta quindi un notevole ausilio ai gestori delle basi di dati pubbliche per la progettazione di specifici interventi mirati al miglioramento della qualità.
- l'ideazione di un processo che consente di integrare i vari contributi sopra illustrati in una strategia per il miglioramento della qualità delle basi di dati toponomastiche che le amministrazioni pubbliche possono seguire.

I principi guida proposti prevedono diverse categorie di utenti.

Le proposte relative ai formati di acquisizione e di interscambio e agli interventi sui processi inter-amministrazioni interessano tutti gli enti della Pubblica Amministrazione che hanno basi di dati in cui vengono acquisiti, memorizzati e scambiati dati toponomastici.

Gli utenti della metodologia di misurazione e miglioramento sono ancora tutte le pubbliche amministrazioni, centrali e locali, che abbiano basi di dati toponomastici. Si anticipa però che la metodologia di misurazione presenta aspetti avanzati che ha senso applicare a basi di dati toponomastiche di un ordine di grandezza di almeno 100.000 record. Infatti, alcune tecniche proposte sono costose ed impegnative in termini di risorse.

1.2 Il Contesto

Il contesto di riferimento dei principi guida è di seguito introdotto mediante una prima definizione dei concetti di dato toponomastico e qualità dei dati (Paragrafo 1.2.1) e, successivamente, mediante il concetto di cooperazione tra le amministrazioni (Paragrafo 1.2.2).

1.2.1 Dati Toponomastici e Qualità dei Dati

Le specifiche proposte riportate, nel presente volume, sono state sviluppate avendo come riferimento la relazione che deve sussistere tra il miglioramento della qualità e lo specifico contesto dei dati toponomastici.

Il contesto in oggetto è definibile in base ai seguenti aspetti: *(i)* la *tipologia delle basi di dati*; *(ii)* i *processi* di acquisizione, trattamento e diffusione dei dati; *(iii)* i *soggetti* (fisici o giuridici) coinvolti in tali processi. In questa parte introduttiva ci si limita a fornire alcuni cenni a tali aspetti che saranno approfonditi nel seguito.

Per quanto riguarda la *tipologia delle basi di dati*, è opportuno distinguere tra *basi di dati toponomastici puri o indirizzi*, che contengono unicamente dati toponomastici e *basi di dati di localizzazione*, che contengono le associazioni tra dati toponomastici e i soggetti fisici o giuridici.

In relazione ai *processi*, è necessario specificare i processi connessi agli eventi di *creazione e/o modifica* di specifici aggregati territoriali (ad esempio la creazione di un comune) e i processi connessi alla *diffusione* degli eventi di creazione e/o modifica. I processi del primo tipo hanno un impatto sulle basi di dati toponomastici puri, mentre i processi del secondo tipo hanno un impatto sia sulle basi di dati toponomastici puri che sulle basi di dati di localizzazione.

Analogamente a quanto appena illustrato, è pertanto necessario trattare in modo separato: i soggetti che *gestiscono* le *basi di dati toponomastici puri* e i soggetti che gestiscono le *basi di dati di localizzazione*.

Le dimensioni della qualità dei dati che è rilevante considerare, al fine di migliorare la qualità nel contesto appena descritto, sono: l'*accuratezza sintattica*, la *completezza*, la *consistenza interna* e l' *aggiornamento*.

L'accuratezza sintattica esprime la vicinanza tra i valori delle componenti di un dato toponomastico e i valori esatti (o ufficiali) delle componenti stesse.

La completezza considera la presenza o l'assenza di valori per ciascuna componente del dato toponomastico.

La consistenza interna valuta il rispetto di regole che legano due o più componenti del dato toponomastico.

L'aggiornamento[1] del dato misura la vicinanza di un valore di una componente del dato toponomastico con il valore corrente (o ufficiale).

1.2.2 La Cooperazione tra le Pubbliche Amministrazioni

I principi guida considerano le amministrazioni pubbliche come parte di un sistema cooperativo. I sistemi di cooperazione (Cooperative Information Systems, CIS) [CIS] coinvolgono organizzazioni indipendenti che decidono di cooperare, su obiettivi comuni, per fornire servizi a valore aggiunto. Lo sviluppo del Sistema Informativo Unitario della Pubblica Amministrazione, secondo l'architettura di cooperazione, è un esempio di CIS.

Nell'ambito della strategia di *e*-Government del Governo Italiano, è prevista la realizzazione, a partire dalle reti esistenti, di una rete telematica a copertura nazionale in grado di interconnettere tutti i sistemi informativi delle amministrazioni e degli enti locali e centrali, e che consenta, in condizioni di sicurezza, lo scambio di servizi applicativi paritetici tra tutte le amministrazioni. In particolare, l'obiettivo è quello di creare un Sistema Informativo Unitario della Pubblica Amministrazione, integrando l'insieme dei differenti sistemi informativi delle singole amministrazioni, autonomi, distribuiti ed eterogenei, secondo una comune *architettura di cooperazione applicativa*. I principi ispiratori dell' architettura di cooperazione sono quelli dell'autonomia e della cooperazione a livello applicativo:

- ogni amministrazione *client* deve essere in grado di scambiare informazioni ed accedere ai servizi applicativi delle altre, in modo trasparente rispetto all'architettura dei sistemi informativi delle amministrazioni *server* (che spesso si configurano come *legacy system*);
- i vincoli imposti ad ogni amministrazione possono essere di natura tecnologica, organizzativa e normativa. Nello stabilire un'architettura di cooperazione comune, è fondamentale rispettare l'autonomia di ciascuna amministrazione e tali vincoli devono dunque essere limitatamente stringenti.

I principi di autonomia e cooperazione applicativa sono alla base del quadro metodologico proposto. In particolare, come descritto nel Capitolo 5, la proposta prevede in maniera esplicita un miglioramento della qualità dei dati toponomastici basato sulla cooperazione tra le amministrazioni.

1.3 Progetti Connessi

In questo paragrafo si introducono brevemente alcuni progetti nazionali ed internazionali in materia di integrazione, diffusione e standard relativi ai dati toponomastici e geografici.

---

[1] Si noti che per aggiornamento non si intende l'*operazione* di aggiornamento del dato, ma la caratteristica di validità temporale del dato nell'istante di osservazione.

Come si è avuto modo di evidenziare, uno dei limiti presenti nell'attuale organizzazione dell'informazione toponomastica, è quello della mancanza di riferimenti ufficiali e condivisi, relativi a dati toponomastici e cartografici. Occorre considerare che, sia in ambito nazionale che in ambito europeo sono stati avviati alcuni importanti progetti aventi per obiettivo proprio la costruzione di database topografici e cartografici di interesse comune.

Il primo progetto che occorre certamente menzionare è il progetto IntesaGIS [IntesaGIS], originato dall'Accordo Integrativo sul Sistema Cartografico di Riferimento, approvato dalla Conferenza Stato-Regioni il 12 ottobre 2000, che coinvolge vari Enti Centrali, gli Istituti aventi compiti in materia cartografica, le Regioni coordinate dalle Regioni capofila e dal Centro Interregionale. I principali obiettivi di tale progetto sono: *(i)* definire un nuovo modello di fruizione delle informazioni cartografiche e topografiche nel rispetto delle autonomie locali; *(ii)* creare e gestire una base di dati centralizzata dei metadati cartografici e topografici esistenti; (*iii*) creare adeguati motori di ricerca; *(iv)* creare un'infrastruttura di identificazione ed autorizzazione degli utenti; *(v)* gestire i metadati cartografici e topografici disponibili presso le Pubbliche Amministrazioni (Comuni, Enti Centrali, Province e Regioni). Ad oggi è stata completata la realizzazione dell'infrastruttura ed alcune procedure per la gestione della base dei metadati e per la georeferenziazione dei dati spaziali.

Il progetto IntesaGIS si inserisce poi nel più ampio progetto europeo denominato INSPIRE (INfrastructure for SPatial InfoRmation in Europe) [INSPIRE], con cui si vuole costruire un sistema europeo integrato per l'accesso all'informazione geografica e spaziale da parte di tutta l'utenza, enti locali, ricercatori e *policy makers*.

Occorre infine menzionare anche che Eurogeographics [Eurogeographics] sta portando avanti il progetto Eurospecs, che si propone nei prossimi anni di identificare specifiche comuni relative ai dati toponomastici.

I progetti descritti considerano i dati toponomastici con riferimento a sistemi cartografici e di georeferenziazione. La principale differenza con l'approccio adottato nella concezione dei principi guida, è nella necessità di identificare, in maniera specifica, unità ecografiche piuttosto che genericamente punti sul territorio. Tale concetto sarà ulteriormente approfondito nel Capitolo 2.

Due progetti importanti in ambito nazionale, che hanno affrontato alcune problematiche connesse ai dati toponomastici, sono: il Repertorio Integrato degli Agenti Economici, nel seguito indicato come RAE, e il Sistema di Accesso ed Interscambio Anagrafico, nel seguito indicato come SAIA.

Il RAE [RAE] ha coinvolto tre importanti enti amministrativi INPS, INAIL e Camere di Commercio, in relazione ai dati relativi alle imprese. Uno degli obiettivi raggiunti dal progetto RAE è stato la costruzione di un archivio integrato delle imprese, per il quale è stato necessario affrontare aspetti relativi al trattamento e alla riconciliazione di indirizzi connessi alle imprese e presenti negli archivi suddetti.

Il progetto SAIA [SAIA] si propone di: *(i)* garantire l'interconnessione dei Comuni e razionalizzare l'interazione tra questi e le Amministrazioni centrali e territoriali in materia di informazione anagrafica; *(ii)* garantire le funzioni di supporto necessarie alla emissione della carta di identità elettronica; *(iii)* garantire la presenza

dell'iscrizione di un cittadino in una sola anagrafe comunale e di eliminare le eventuali duplicazioni d'iscrizione. L'architettura del SAIA è basata sull'Indice di Riferimento Nazionale (INA) che consente il collegamento logico virtuale delle anagrafi comunali per il reperimento su base nazionale degli indirizzi di residenza dei cittadini.

Nell'elaborazione dei principi guida sono stati considerati i risultati di entrambi i progetti, con particolare riferimento all'interscambio degli indirizzi per la localizzazione dei soggetti fisici e giuridici.

1.4 Guida alla Lettura del Volume

L'obiettivo di questo paragrafo è di fornire una guida al contenuto del presente volume.

Si possono individuare tre diverse viste per la lettura del volume, corrispondenti ad un interesse primario rispettivamente in :

- *caratterizzazione della qualità dei dati toponomastici*. Consiste nella definizione delle dimensioni della qualità associate al dato toponomastico (Paragrafo 2.2). Inoltre, una metodologia per associare delle misure di qualità a ciascuna delle dimensioni definite è descritta nel Capitolo 3. Tale capitolo può essere concettualmente suddiviso in due parti. Nella prima parte (Paragrafo 3.1), è definito un insieme di indicatori atti a descrivere quantitativamente alcune dimensioni di qualità. Nella seconda parte (Paragrafo 3.2) si affrontano i problemi di implementazione della misura degli indicatori definiti nella prima parte. Inoltre, si illustra un metodo di misurazione più complesso che consente di identificare sottoinsiemi dei record memorizzati nelle basi di dati toponomastiche tali da presentare caratteristiche di qualità differenti;
- *miglioramento della qualità dei dati toponomastici*. Si sviluppa secondo due direttive che corrispondono al miglioramento interno alle amministrazioni (Capitolo 4) ed al miglioramento basato sulla cooperazione tra le amministrazioni (Capitolo 5). Il capitolo 4 contiene i principi guida per il miglioramento dell'accuratezza sintattica e della completezza, definendo un formato standard di acquisizione dei dati toponomastici. Tale formato riguarda l'interazione tra il cittadino e una singola amministrazione e può essere cartaceo ed elettronico. Relativamente al caso di acquisizione elettronica sono stati definiti dei dizionari per le componenti Comune, Provincia, Località, CAP e Denominazione Urbanistica Generica o DUG. Tali dizionari, allegati in Appendice [2], consentono di effettuare i necessari controlli sui dati acquisiti, così garantendo l'accuratezza e la completezza dell'informazione immessa nelle basi di dati delle singole amministrazioni. Il Capitolo 5 contiene i principi guida per il miglioramento dell'accuratezza sintattica e dell'aggiornamento, definendo un formato standard di

[2] Le appendici del volume, ad eccezione del glossario, della normativa e del documento preliminare al capitolo 3, sono disponibili in formato elettronico presso il Servizio PSM dell'ISTAT.

interscambio dei dati toponomastici (Paragrafo 5.1) e un insieme di indicazioni per la reingegnerizzazione dei flussi informativi, al fine di garantire l'allineamento delle basi di dati toponomastici (Paragrafo 5.2). In particolare, in merito ai flussi informativi considerati, sono stati approfonditi i flussi connessi agli eventi di creazione e/o modifica di componenti del dato toponomastico puro;

– *approccio globale alla gestione e al miglioramento della qualità dei dati toponomastici*. Una chiave unitaria di lettura delle singole metodologie proposte per la misurazione ed il miglioramento della qualità dei dati toponomastici è presentata nel Paragrafo 2.3. In tale paragrafo, si suggerisce alle amministrazioni di seguire un processo integrato per misurare e migliorare la qualità dei propri dati toponomastici. Tale processo consta di fasi, e ciascuna fase richiede l'applicazione delle metodologie e delle tecniche specifiche suggerite nei capitoli 3,4,5.

2 Quadro Metodologico di Riferimento

Il presente capitolo è dedicato alla descrizione del quadro metodologico di riferimento, che fornisce le definizioni dei concetti base per una lettura unitaria dei principi guida che consentono alle amministrazioni pubbliche di sviluppare azioni coordinate per la misurazione ed il miglioramento dei propri dati toponomastici.

Il capitolo consta di tre paragrafi. Nel paragrafo 2.1, sono fornite alcune definizioni preliminari e, a partire da un'analisi empirica di importanti basi di dati di enti pubblici, vengono considerate le componenti che caratterizzano il dato toponomastico. Nel paragrafo 2.2, sono definite ed esemplificate le dimensioni di qualità dei dati ritenute rilevanti per il contesto in esame. Nel paragrafo 2.3 è descritto il processo integrato suggerito alle amministrazioni pubbliche, per la misurazione ed il miglioramento dei dati toponomastici.

2.1 Analisi e Classificazione Preliminare dei Dati Toponomastici

2.1.1 Le Componenti dei Dati Toponomastici

Nel presente paragrafo si individuano e si analizzano le componenti del dato toponomastico che consentono l'individuazione sul territorio dei soggetti fisici e giuridici.

La localizzazione esatta di un *luogo fisico* è data dalle coordinate geografiche espresse in un sistema di riferimento noto.

Ad un *luogo fisico* può essere associata un' *unità ecografica*, che può essere definita secondo differenti livelli:

– *Livello 1-unità ecografica semplice*. E' costituita da: *(i)* una abitazione, cioè uno o più vani funzionalmente destinati alla vita delle persone, *(ii)* oppure da un

esercizio, cioè uno o più vani funzionalmente destinati allo svolgimento di una qualsiasi attività, ad esempio di tipo economico;

- *Livello 2-unità ecografica complessa fabbricato*. E' costituita da un insieme di unità ecografiche semplici, il cui accesso o i cui accessi esterni sono contraddistinti da numeri civici;
- *Livello 3-altre unità ecografiche complesse di livello superiore*.Un esempio sono gli isolati, definiti come insiemi di fabbricati normalmente delimitati da spazi destinati alla viabilità.

Nell'ambito del progetto, si sono considerate le unità ecografiche del livello due, nel seguito indicate genericamente come *unità ecografiche*. In particolare, ci si è concentrati sulla problematica dell'identificazione degli accessi a tali unità. Tale scelta non consente la puntuale identificazione di unità ecografiche semplici (qualora aggregate in fabbricati). L'identificazione puntuale delle unità ecografiche semplici rientra in particolari processi come, ad esempio, i processi amministrativi di interesse per il catasto del Ministero dell'Economia e delle Finanze. Invece, la scelta di concentrarsi sugli accessi ai fabbricati è dettata dall'esigenza di coprire la maggioranza dei processi amministrativi, che non sono interessati al livello di dettaglio dell'identificazione delle unità ecografiche semplici, o di unità ecografiche complesse di livello superiore. Infatti, come sarà esplicitato a breve, l'identificazione dell'accesso ai fabbricati consente la localizzazione dei soggetti fisici e giuridici, come memorizzati nella maggior parte delle basi di dati toponomastici della pubblica amministrazione.

L'accesso alle unità ecografiche potrebbe essere identificato dalle coordinate geografiche del numero civico. Invece, comunemente, tali accessi sono identificati in modo simbolico, ed è questa la scelta di identificazione che si è approfondita nei principi guida.

In particolare, si definisce come *dato toponomastico puro (o dato toponomastico o indirizzo)* una sequenza di elementi simbolici, che identificano in modo univoco l'accesso alle unità ecografiche definite come sopra.

Il *dato toponomastico di localizzazione* è definito come un'associazione tra il dato toponomastico puro e un soggetto fisico o giuridico. Diversi soggetti fisici o giuridici possono condividere il medesimo accesso ad un'unità ecografica, e quindi lo stesso indirizzo.

Gli *elementi simbolici* che costituiscono il dato toponomastico puro, corrispondono ad oggetti di tipo amministrativo e/o topografico universalmente noti. Sono *elementi simbolici* i nomi di una località geografica (ad esempio, di un comune), i nomi di una via o piazza o simili, i numeri civici, il codice di avviamento postale.

Nel seguito si identificano le entità costituenti il dato toponomastico a partire dalle proprietà strutturali dello stesso. In particolare, un dato toponomastico è una struttura gerarchicamente complessa, composta da più livelli. Una prima suddivisione di massima identifica due livelli: il primo livello è costituito da una *zona di territorio*; il secondo livello è costituito da un *punto sul territorio* all'interno di una zona.

Entrambi questi livelli sono costituiti da entità elementari.

Per quanto riguarda una *zona del territorio*, non esiste allo stato attuale una definizione condivisa sulle entità elementari che la costituiscono. Nella sua accezione più ampia e diffusa, la zona del territorio corrisponde ad un *comune*. Si noti che per identificare in modo univoco un comune, è necessario specificare la *provincia* di appartenenza (in province diverse possono esistere comuni con gli stessi nomi). E' tuttavia possibile che molte volte con il termine zona di territorio si faccia riferimento ad una *località*, che può essere una località postale (cfr. Appendice 1: Glossario) o un aggregato territoriale sub-comunale, come la località abitata ISTAT (cfr. Glossario). Mentre le entità introdotte sono in relazione gerarchica, un'ulteriore entità che compare con riferimento alla zona di territorio e che invece non gode di tale relazione è il *Codice di Avviamento Postale* (CAP). La suddivisione territoriale prevista dai CAP tipicamente identifica aggregazioni di più comuni, zone intra-comunali.

Per quanto riguarda un *punto sul territorio*, questo è individuabile come associazione di due entità: l'*area di circolazione* e il *numero civico*. L'area di circolazione, in un territorio dotato di regolare rete stradale, è l'elemento lineare o areale di qualsiasi forma o misura destinato alla viabilità. Le aree di circolazione possono essere di tipo diverso, ad esempio via, strada, corso, viale, vicolo, calle, salita, piazza, piazzale, largo, campiello e simili. La tipologia dell'area di circolazione è specificata dalla *Denominazione Urbanistica Generica (DUG).* Ai fini della sua esatta individuazione, ciascuna area di circolazione, oltre che dalla DUG, è contraddistinta da una propria *denominazione*.

Il numero civico è la componente che individua l'accesso ad un'unità ecografica nell'ambito di un'area di circolazione.

In sintesi, le entità elementari che compaiono in ciascuno dei due livelli descritti sono rappresentate in Figura 2.1. In figura, sono anche specificate le relazioni di tipo gerarchico tra le entità elementari all'interno di ogni livello.

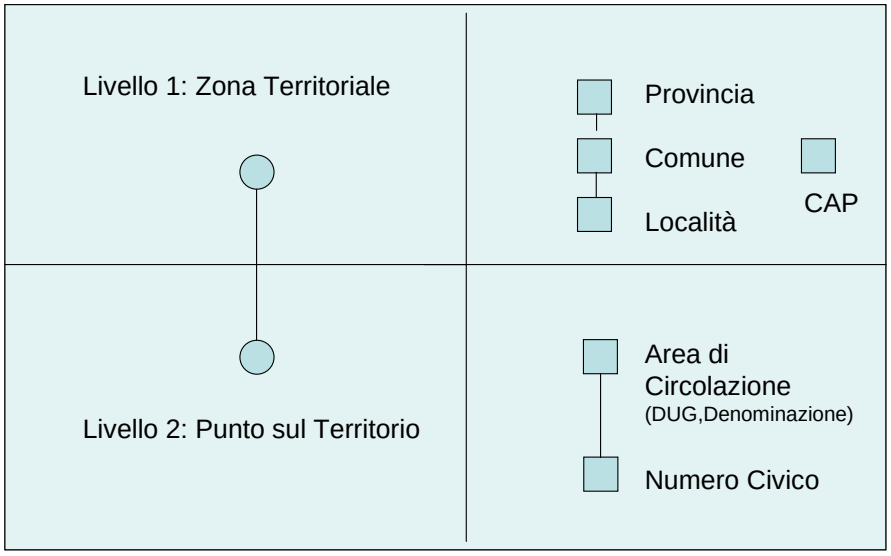


Figura 2.1: La struttura a livelli del dato toponomastico e le sue componenti elementari.

2.1.2 La Rappresentazione del Dato Toponomastico in alcune Basi di Dati Nazionali

Al fine di verificare la corrispondenza tra la struttura del dato toponomastico, descritta nel paragrafo precedente, e la sua rappresentazione nelle basi di dati toponomastici, è stata fatta un'analisi di come tali dati sono rappresentati nelle basi di dati di alcuni enti pubblici.

In particolare, sono state considerate due basi di dati toponomastici puri, appartenenti ad ISTAT e Poste Italiane. Inoltre, sono state considerate quattro basi di dati di localizzazione di soggetti fisici e/o giuridici; queste appartengono ai seguenti enti: SEAT, Camere di Commercio, Agenzia delle Entrate del Ministero della Economia e delle Finanze e INPS.

Nella Tabella 2.1, che segue, si riportano sinteticamente i risultati di tale analisi.

Nella base di dati dell' Istat, i dati toponomastici sono quelli raccolti in occasione dei censimenti generali del 1991. Tali dati si riferiscono agli itinerari delle sezioni di censimento (cfr. Glossario) dei singoli comuni, a partire dai quali è stato derivato lo Stradario Nazionale Istat. Si noti come tale archivio non contiene il riferimento puntuale di tutti i numeri civici, ma solo gli intervalli che appartengono a ciascun arco di strada (cfr. Glossario). Inoltre è presente una tabella delle DUG molto completa e articolata che comprende la DUG e altre parole di complemento (a, di, alla, ecc.). Una caratteristica importante della base di dati dell' Istat è l'associazione del dato toponomastico al codice della sezione di censimento nella quale ricade (geocodifica) e/o alle coordinate geografiche che identificano la sezione stessa. In sintesi, la struttura del dato toponomastico memorizzato nell'archivio dell'ISTAT comprende i seguenti elementi: provincia, comune, CAP, DUG e denominazione dell'area di circolazione, numero civico (rappresentato come intervallo dei numeri civici per arco di strada). Nell'archivio Istat vengono, inoltre, memorizzate anche le "dizioni alternative" per comune e località, al fine di gestire le doppie denominazioni per i comuni a doppia lingua o per quelle denominazioni riconosciute dalla popolazione locale con una dizione alternativa alla denominazione ufficiale, o ancora vecchie denominazioni che cambiano nome. Nella base di dati di Poste Italiane, sono raccolte le informazioni relative allo stradario nazionale e all'organizzazione del recapito postale. La struttura del dato toponomastico risulta più articolata. La DUG è suddivisa in due componenti: la DUG vera e propria e il Complemento della DUG, che contiene le parole di complemento alla DUG (a, di, alla, ecc.). Anche la Denominazione dell'area di circolazione è suddivisa in due componenti distinte: la denominazione vera e propria e il complemento alla denominazione, che corrisponde tipicamente a titoli onorifici per i nomi propri, che costituiscono la denominazione dell'area di circolazione (ad esempio, generale). Come nel caso dell'archivio Istat, anche l'archivio di Poste Italiane memorizza informazioni che supportano il bilinguismo.

Nella base dati SEAT Pagine Gialle, relativa agli indirizzi delle utenze telefoniche, il dato toponomastico è articolato nelle seguenti componenti: comune, località/frazione, DUG, denominazione dell'area di circolazione, numero civico.

Componenti del dato toponomastico	Basi di dati toponomastici puri		Basi di dati di localizzazione			
	Istat	Poste Italiane	SEAT	Camere di commercio	Agenzia delle Entrate	INPS
Provincia	*	*		*	xx	*
Comune	*	*	*	*	xx	*
Località/frazione		*	*	*		
CAP	*	*		*	*	*
Area di circolazione					*	
DUG	*	*	*	*		*
Compl.alla DUG		*				
Denominazione	*	*	*	*		*
Compl.alla Denominazione		*				
Numero civico	X	*	*	*		*

Tabella 2.1: Componenti del dato toponomastico nelle basi di dati di alcuni Enti.
Note:
*: presenza della componente nella base di dati.
x : Sono memorizzati gli intervalli di numeri civici (intervalli “da”, “a” di un arco di strada).
xx : Si ricavano decodificando il codice catastale del comune.

Il campo località/frazione è considerato facoltativo. La località/frazione è compresa all’interno del comune. E’ inoltre presente in SEAT un archivio delle località contenente circa 30.000 occorrenze, ricavate dagli indirizzi acquisiti. Le parole di complemento alla DUG sono considerate parte della DUG stessa. E’ inoltre presente un campo “indicazioni in parentesi” per la specifica di informazioni aggiuntive soprattutto in assenza di numero civico.

Nella base di dati delle Camere di Commercio che contiene gli indirizzi delle unità locali delle imprese, il dato toponomastico è articolato nelle seguenti componenti: provincia (espressa come sigla), comune, DUG, denominazione dell’area di circolazione, numero civico, CAP, frazione, altre indicazioni sull’ indirizzo. Il campo “altre indicazioni indirizzo” viene utilizzato quando si ritiene che le informazioni presenti nei campi: DUG, denominazione dell’area di circolazione e numero civico, non siano sufficienti ad individuare esattamente un indirizzo.

Nella base dati dell’Agenzia delle Entrate, riferita ai soggetti fisici e alle persone giuridiche titolari di partita IVA, sono riportati gli indirizzi della sede legale e del domicilio fiscale delle imprese. Entrambi gli indirizzi sono costituiti dai seguenti elementi: codice catastale del comune, area di circolazione e numero civico (in un unico campo), CAP. La componente area di circolazione contiene oltre al numero civico la DUG e la denominazione dell’area di circolazione. Il codice catastale del

comune permette, attraverso un apposita tabella di decodifica, di ricavare la provincia e il comune corrispondenti al codice.

L'ultima base di dati di localizzazione esaminata è l'archivio anagrafico dell'INPS, che contiene gli indirizzi delle unità contributive. Il dato toponomastico è memorizzato mediante le seguenti componenti: provincia, comune, DUG, denominazione dell'area di circolazione, numero civico, CAP.

Dall'analisi della struttura del dato toponomastico nelle basi di dati esaminate si può rilevare che le componenti ricorrenti sono: la provincia, il comune, l'area di circolazione, composta da DUG e denominazione e il numero civico. Ne consegue che una qualsiasi base di dati toponomastici deve contenere almeno questo insieme di componenti, che può essere considerato un insieme minimo di riferimento. A partire da tale insieme sono stati definiti dei formati standard per l'acquisizione e l'interscambio dei dati toponomastici, per i quali si rimanda ai capitoli 4 (per quanto riguarda lo standard di acquisizione) e 5 (per quanto riguarda lo standard di interscambio).

2.2 Le Dimensioni della Qualità dei Dati Toponomastici

La definizione delle dimensioni della qualità dei dati è un tema che è stato sviluppato in diversi contesti scientifici. In particolare, sia in ambito statistico che informatico esiste una letteratura specifica relativa all'insieme delle dimensioni di qualità che meglio caratterizza il concetto di qualità dei dati.

Nel campo statistico gli studi effettuati negli istituti di ricerca hanno condotto all'identificazione di diverse dimensioni caratterizzanti la qualità dei dati. Eurostat definisce sette dimensioni: *rilevanza* dell'informazione, *accuratezza* delle stime, *tempestività* o *puntualità* nella diffusione dei risultati, *accessibilità* o *chiarezza* delle informazioni, *confrontabilità* delle informazioni nel tempo e nello spazio, *coerenza* nell'ambito della stessa fonte di riferimento o tra più fonti, *completezza* del quadro informativo del dominio di interesse ([Eurostat1999], [Eurostat2000] [Eurostat2001A], [Eurostat2001B]). Nel contesto di Statistics Canada viene adottato un criterio di valutazione della qualità secondo sei dimensioni: *rilevanza, accuratezza, tempestività, accessibilità, confrontabilità, coerenza* ([BrackstoneGordon1999], [StatisticsCanada1998]).

Anche nel campo informatico sono state effettuate diverse proposte di dimensioni. Una prima proposta [WangStrong1996] prevede di individuare quattro categorie per le dimensioni di qualità: intrinseca, dipendente dal contesto, relativa alla rappresentazione, relativa alla accessibilità. La tassonomia proposta in [Redman1996] include più di venti dimensioni di qualità dei dati, classificate in tre categorie: vista concettuale, vista dei valori, vista del formato. In generale, le dimensioni di qualità più utilizzate e note in ambito informatico sono: accuratezza, completezza, consistenza e aggiornamento. Una rassegna delle proposte di dimensioni di qualità dei dati in ambito informatico è fornita in [Wang1995].

Inoltre, in [AIPA2002], un insieme di dimensioni della qualità dei dati sono definite, nell'ambito delle linee guida per l'accesso, la comunicazione e la diffusione dei dati

pubblici. Tale insieme include: sicurezza, usabilità, esattezza, accuratezza, completezza, consistenza, tempestività e pertinenza.

Nel presente progetto, si è adottato un insieme di dimensioni di qualità che caratterizza in maniera specifica la qualità dei dati toponomastici, piuttosto che la qualità dei dati in generale. Le dimensioni di qualità considerate sono: l'*accuratezza sintattica*, la *consistenza interna*, la *completezza* e l'*aggiornamento.* Ciascuna di queste dimensioni è infatti rilevante per la caratterizzazione della qualità dei dati toponomastici: l'accuratezza sintattica consente di associare un grado di correttezza ai dati toponomastici; la consistenza interna cattura le dipendenze tra i campi (ad esempio, CAP e comune sono campi inter-dipendenti); la completezza è particolarmente importante per caratterizzare l'assenza di valori su quei dati necessari all'identificazione univoca di un dato toponomastico (ad esempio, l'assenza del campo provincia può non consentire l'identificazione di un comune, essendo ammissibili comuni con lo stesso nome in province diverse); l'aggiornamento è importante per rappresentare la variabilità dei dati toponomastici nel tempo (ad esempio, è importante poter individuare, date due Denominazioni per la medesima area di circolazione, quale è quella corrente).

Le dimensioni della qualità considerate si rapportano alla letteratura, in ambito statistico ed informatico, come segue:

con riferimento all'insieme delle dimensioni definite come principali in ambito informatico, le dimensioni adottate sono costituite da tutte e sole le quattro dimensioni incluse nell' insieme preso in esame;

con riferimento alle dimensioni definito in ambito europeo, sono escluse dall'insieme considerato le dimensioni della qualità relative alla rilevanza e alla accessibilità. Infatti, queste ultime, pur non essendo esplicitamente definite come dimensioni misurabili, hanno comunque, guidato alcune scelte progettuali. Ad esempio, la scelta di un insieme minimo di componenti del dato toponomastico, è stata guidata dalla rilevanza. Si noti, inoltre, che la dimensione dell'aggiornamento unifica le due dimensioni statistiche di tempestività e confrontabilità.

Con riferimento alle dimensioni considerate in [AIPA2002], si noti che le dimensioni selezionate nel presente lavoro, ne costituiscono un sottoinsieme. In merito alle dimensioni escluse, che riguardano la pertinenza e l'usabilità, vale quanto detto per la rilevanza e l'accessibilità nel caso delle qualità statistiche. L'esattezza, come grado di accuratezza massimo, è riconducibile all'accuratezza. La sicurezza non è stata considerata nell'ambito delle dimensioni misurabili, ma alcuni aspetti da essa implicati sono presi in esame nella definizione dei flussi di aggiornamento, illustrati nel Capitolo 5. Ad esempio, la disponibilità e l'integrità dell'informazione toponomastica sono considerate mediante la gestione dei flussi di aggiornamento da parte dei *Data Steward* delle varie componenti del dato toponomastico.

Nel Capitolo 3 del presente volume, sarà definito un insieme di indicatori di qualità, che si riferiscono alle singole componenti del dato toponomastico identificate nel paragrafo 2.1.2. In particolare, saranno definiti indicatori per la misura del grado di accuratezza sintattica, di completezza e di consistenza interna di dati toponomastici, presenti in una base di dati amministrativa. Nel Capitolo 5, invece, si forniranno

indicazioni metodologiche per il miglioramento dell'aggiornamento, basate sul ridisegno dei processi inter-amministrazioni.

Nel seguito del presente paragrafo, sono fornite le definizioni di riferimento per ciascuna delle dimensioni di qualità considerate nei principi guida.

2.2.1 Accuratezza Sintattica

L'accuratezza, in generale, fa riferimento al grado di prossimità tra uno specifico dato e un valore preso a riferimento. E' misurata mediante indici di *distanza* tra il valore assunto dal dato osservato e quello preso come riferimento. Nel caso dei dati toponomastici, l'accuratezza misura la distanza tra il valore della componente toponomastica in esame e il valore ufficiale della denominazione o della codifica di tale componente. Nei principi guida, l'accuratezza è denotata come *accuratezza sintattica*.

I principali strumenti che si propongono alle amministrazioni per migliorare l'accuratezza sintattica sono:

- la definizione di uno standard per l'acquisizione dei dati toponomastici;
- la proposta di un insieme di dizionari di riferimento per le componenti del dato toponomastico;
- la proposta di tecniche di valutazione dell' accuratezza sintattica, che può essere misurata attraverso la costruzione di indicatori specifici di qualità, ottenuti per mezzo di adeguate tecniche statistiche (come, ad esempio, le tecniche di regressione non parametrica [Breiman1984]). Tali indicatori di qualità possono essere definiti per una componente del dato toponomastico, per più componenti, o per l'intera base dati come misura sintetica di accuratezza sintattica;
- l'interscambio automatico, abilitato dalla definizione di uno standard. Tale interscambio, proposto nell'ambito del ridisegno dei processi amministrativi, supporta il miglioramento dell'accuratezza sintattica consentendo verifiche dinamiche della conformità ai dizionari di riferimento per i dati scambiati.

2.2.2 Consistenza Interna

La consistenza interna tra due o più componenti del dato toponomastico verifica il rispetto di una regola semantica che lega tali componenti.

Un errore di consistenza interna può essere evidenziato, ad esempio, da una delle seguenti circostanze:

- non coerenza di una denominazione con una codifica, oppure non congruenza di due codifiche diverse che fanno riferimento a un medesimo dato toponomastico; ad esempio, in una medesima base di dati, il codice comune dell'ISTAT può non essere coerente con il codice comune del Ministero delle Finanze;
- non consistenza delle componenti del dato toponomastico in relazione gerarchica; ad esempio, un comune ed una provincia possono essere erroneamente associati, perché il comune non appartiene a quella provincia.

Il miglioramento della consistenza interna può essere conseguito adottando i medesimi strumenti elencati per il miglioramento dell'accuratezza sintattica.

2.2.3 Completezza

La completezza può essere considerata sotto diversi aspetti:

- completezza rispetto all'informazione elementare, che implica la presenza o l'assenza del valore per una componente del dato toponomastico;
- completezza rispetto ad una popolazione teorica di riferimento. Questa dimensione della qualità misura la copertura di una base di dati amministrativa rispetto alla *popolazione teorica* di unità che, in base alla normativa esistente, dovrebbe essere riportata nella base di dati.

Poiché le misure di completezza di quest'ultimo tipo sono difficilmente generalizzabili, in quanto collegate alle specifiche normative, che individuano la popolazione dei soggetti di competenza di ciascun ente amministrativo, nel seguito il termine completezza verrà utilizzato esclusivamente con riferimento alla completezza dell'informazione elementare.

Per il miglioramento della completezza, si possono adottare i medesimi strumenti elencati per il miglioramento dell'accuratezza sintattica e della consistenza interna.

2.2.4 Aggiornamento

L'aggiornamento consente di caratterizzare la variabilità nel tempo dei dati toponomastici e di verificare se il dato memorizzato nell'archivio amministrativo corrisponde a quello *valido* nell'istante di valutazione.

Si noti che si ha un errore di accuratezza se, ad esempio, una componente di un indirizzo di residenza è errata, ad esempio numero civico 35 anziché 85; invece, si ha un dato non aggiornato se la persona di cui si considera la residenza ha abitato prima al civico 35 e poi al civico 85. E' importante capire la natura dell'errore, per poter intraprendere azioni di miglioramento opportune. Nell'esempio, per l'errore di accuratezza si dovrà agire sul processo di *data entry* dell'informazione toponomastica; nel caso di dato non aggiornato, invece, bisognerà individuare la fonte primaria di aggiornamento di quel dato e garantire la propagazione degli aggiornamenti alle altre basi di dati che memorizzano il dato stesso.

Nel presente lavoro, sono diffusamente trattate indicazioni metodologiche per il miglioramento dell'aggiornamento, individuando i necessari processi di diffusione per ciascuna componente aggiornata del dato toponomastico (vedi Capitolo 5).

2.3 Il Processo Integrato per la Misurazione ed il Miglioramento della Qualità dei Dati Toponomastici

L'obiettivo di questo paragrafo è di illustrare il processo che si suggerisce alle amministrazioni pubbliche per attuare la misurazione e il miglioramento della qualità dei propri dati toponomastici. Come vedremo, tale processo consta di fasi, e ciascuna fase richiede l'applicazione di una metodologia e di tecniche specifiche.

Migliorare la qualità dei dati è un'attività complessa che ha richiesto l'ideazione e la proposta di processi metodologici più o meno articolati (e.g., [Redman1996], [Shankaranayan2000]). Tali processi hanno principalmente riguardato azioni di miglioramento interno a singole organizzazioni, laddove, invece, il presente lavoro si è anche concentrato sull'insieme delle pubbliche amministrazioni e sulle caratteristiche di cooperazione tra le stesse, già esistenti o previste dai programmi governativi. Tuttavia, il processo proposto si basa su diverse caratteristiche e soluzioni suggerite nel corso dell'ultimo decennio da esperti in qualità dei dati sia in ambito accademico che in ambito industriale. In particolare, come base del processo di misurazione e miglioramento che si suggerisce alle pubbliche amministrazioni, si sono adottati i principi di seguito elencati:

il principio di miglioramento della qualità dei dati basato sulla riprogettazione dei flussi informativi, al fine di garantire un miglioramento sul lungo termine. Secondo tale principio è necessario individuare le cause degli errori nei dati che sono insite nei processi di manipolazione dei dati stessi e rimuovere tali cause riprogettando i processi in maniera idonea a garantire la qualità dei dati;

il principio secondo cui qualsiasi intervento di miglioramento della qualità di grandi volumi di dati, necessita di una fase di misurazione che possa orientare le azioni di miglioramento effettive. L'orientamento ha come scopo da un lato la valutazione dell'entità degli errori nei dati; dall'altro, ha l'obiettivo di localizzare gli errori su specifiche tipologie di dato e, quindi, guidare in maniera opportuna le azioni di ridisegno dei processi coinvolti nella manipolazione di tali tipologie di dati.

2.3.1 Il Processo

In questo paragrafo è dettagliato il processo di misurazione e di miglioramento della qualità dei dati toponomastici suggerito alle pubbliche amministrazioni.

Il processo è rappresentato in Figura 2.2. Sulla base dei principi metodologici suggeriti dalla letteratura del settore, e brevemente sopra descritti, si è scelto di prevedere, come fasi del processo di miglioramento: la misurazione della qualità dei dati toponomastici e il miglioramento basato sui processi interni alle amministrazioni.

Inoltre, si è introdotta una fase ulteriore di miglioramento basato sulla cooperazione, che in maniera esplicita considera il contesto di cooperazione che caratterizza le pubbliche amministrazioni italiane.

La fase di misurazione è preliminare alle due fasi di miglioramento interno e miglioramento basato sulla cooperazione. Essa fornisce anche le informazioni di ingresso necessarie all'esecuzione delle successive fasi. Le principali attività da condurre per ciascuna fase sono sinteticamente rappresentate in corrispondenza alle fasi stesse (vedi Figura 2.2).

Nei paragrafi seguenti si fornirà una breve descrizione di ciascuna fase e delle interazioni tra le tre fasi, in termini di successione temporale e prodotti in ingresso ed uscita. Il dettaglio di ciascuna fase sarà invece presentato nei capitoli successivi. In particolare a ciascuna delle tre fasi sarà dedicato un capitolo in cui metodologie e tecniche per conseguire gli obiettivi delle fasi stesse saranno diffusamente descritte.

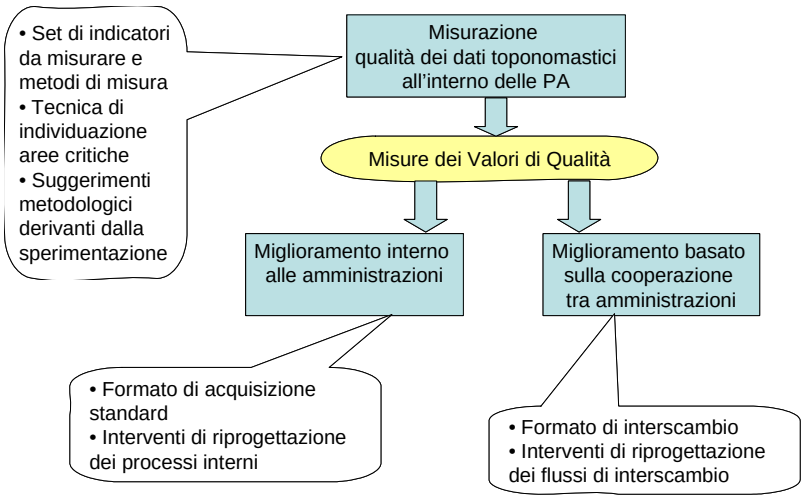


Figura 2.2: Il processo di miglioramento della qualità dei dati toponomastici suggerito alle pubbliche amministrazioni.

2.3.2 La Fase di Misurazione

Nella fase di misurazione della qualità dei dati, si propone alle amministrazioni di adottare tecniche di misura della qualità dei dati toponomastici memorizzati nelle proprie basi di dati. In particolare, si propone alle amministrazioni di condurre la fase di misurazione adottando un approccio base ed eventualmente un approccio avanzato.

L'approccio base consiste nel misurare sulle proprie basi di dati toponomastici, un insieme di indicatori di qualità specificamente proposto nel Paragrafo 3.1.2.

L'approccio avanzato richiede invece l'applicazione di una tecnica che consente l'individuazione di zone del territorio in cui gli errori sintattici sono maggiormente localizzati, dette *aree critiche*. Questo secondo approccio è da adottarsi solo nel caso di basi di dati toponomastiche aventi un ordine di grandezza elevato (almeno 100.000 record).

La fase di misurazione della qualità dei dati toponomastici ha un duplice obiettivo.

In primo luogo, ha lo scopo di orientare le azioni di miglioramento che saranno condotte nella fase di miglioramento interno.

In secondo luogo, ha lo scopo di definire un insieme di valori di qualità per le basi di dati delle amministrazioni pubbliche che potranno utilmente essere esportati alle altre amministrazioni, fornendo così una sorta di autocertificazione della qualità dei propri dati.

La fase metodologica di misurazione della qualità dei dati toponomastici è descritta in dettaglio nel Capitolo 3.

2.3.3 La Fase di Miglioramento Interno

La fase di miglioramento interno è composta da due principali azioni che le amministrazioni possono intraprendere al loro interno, al fine di migliorare la qualità dei propri dati toponomastici.

In primo luogo, sarebbe opportuno l'utilizzo di un formato di acquisizione standard per i dati toponomastici. Uno dei problemi principali dell'informazione toponomastica, infatti, è l'assenza di uno standard che definisca le componenti del dato toponomastico, le condizioni di obbligatorietà o meno di tali componenti e il formato di rappresentazione delle stesse. Per tale motivo, sono stati proposti due formati standard connessi al dato toponomastico: un formato di acquisizione ed un formato di interscambio. Il formato di acquisizione è stato proposto per la relazione tra amministrazione e cittadino. Si è prevista sia la possibilità di acquisizione manuale che di acquisizione automatica, specializzando quest'ultima anche al caso dell'acquisizione via Web. Il formato di interscambio è invece parte della fase di miglioramento basato sulla cooperazione che sarà illustrata nel Capitolo 5.

In secondo luogo, la fase di miglioramento interno prevede l'utilizzo delle informazioni derivate dalla fase precedente di misurazione. In particolare, la misurazione della qualità può essere condotta mediante una serie di indicatori, relativi ad alcune dimensioni della qualità dei dati, e mediante tecniche più avanzate per l'individuazione di zone in cui l'errore è prevalentemente localizzato, anche dette aree critiche. Nell'ambito dei suggerimenti metodologici alle amministrazioni per migliorare internamente la qualità dei dati di propria competenza, si propongono i seguenti interventi:

- utilizzo delle misure di qualità e dei risultati derivanti dalla tecnica di individuazione delle aree critiche per orientare interventi di miglioramento di quei processi interni che manipolano i dati maggiormente affetti da errori;
- utilizzo dei risultati generali derivati dalla misurazione degli indicatori e dalla tecnica della individuazione delle aree critiche sui tre archivi nazionali considerati (vedi Paragrafo 3.2). Anche in questo caso, in maniera complementare alle indicazioni derivanti dalla sperimentazione sulle proprie basi di dati, gli interventi di miglioramento possono essere focalizzati alle aree con dati aventi una maggiore concentrazione di errori.

La fase metodologica di miglioramento interno della qualità dei dati toponomastici è descritta nel dettaglio in Capitolo 4.

2.3.4 La Fase di Miglioramento Basato sulla Cooperazione

La fase di miglioramento della qualità dei dati toponomastici basata sulla cooperazione tra le pubbliche amministrazioni prevede due interventi principali, di seguito elencati.

Il primo intervento consiste nell'adozione di un formato di interscambio standard nell'ambito dei flussi inter-amministrazioni. Si è già accennato all'importanza che riveste la standardizzazione delle componenti del dato toponomastico, sia in merito all'identificazione delle componenti costitutive del dato toponomastico che in merito alla rappresentazione delle stesse. Nell'ambito del presente lavoro, è stata studiata ed elaborata una proposta di formato di interscambio dei dati toponomastici. Tale formato standard ha l'ulteriore caratteristica di supportare l'interoperabilità nello scambio automatico dell'informazione, proponendo che i dati toponomastici vengano scambiati tra le amministrazioni mediante documenti XML [XML2004]. La definizione del formato di interscambio è pertanto avvenuta mediante un linguaggio

per la definizione di schemi XML, in particolare XML Schema ([XMLSchema1], [XMLSchema2]).

Il secondo intervento prevede una riprogettazione dei flussi informativi di scambio dei dati toponomastici, al fine di garantirne l'aggiornamento. E' stato proposto uno schema generale per la riprogettazione dei flussi, che interessa sia le pubbliche amministrazioni, che memorizzano il dato toponomastico puro (ad esempio il catasto), sia le pubbliche amministrazioni che memorizzano dati toponomastici per la localizzazione di soggetti fisici o giuridici (ad esempio, l'anagrafe tributaria del Ministero dell'Economia e delle Finanze).

La fase metodologica di miglioramento della qualità dei dati toponomastici basato sulla cooperazione è descritta nel dettaglio nel Capitolo 5.

3 Misurazione della Qualità dei Dati Toponomastici

Il presente capitolo consta di due parti: la prima parte (Paragrafo 3.1) è finalizzata a definire e formalizzare il quadro concettuale di riferimento, necessario per misurare la qualità dei dati toponomastici negli archivi della pubblica amministrazione; la seconda parte (Paragrafo 3.2) implementa la metodologia proposta su tre rilevanti archivi della pubblica amministrazione. I risultati dell'applicazione sono di notevole interesse, poiché: *(i)* è stato possibile superare le complessità connesse alla misurazione della qualità in situazioni applicative differenti, mettendo a punto algoritmi generalizzati (forniti nel testo o in appendice elettronica); *(ii)* i valori corrispondenti alle misure di qualità riscontrati sui tre archivi costituiscono un utile termine di confronto nella misurazione della qualità di altri archivi pubblici..

3.1 Quadro Concettuale di Riferimento per la Misurazione

Dal punto di vista teorico, nel seguito sarà definito il concetto di *accuratezza sintattica di un indirizzo*, scomponendolo nelle sue tre principali dimensioni, indicate come *ammissibilità*, *completezza* e *consistenza* (Paragrafo 3.1.1). Inoltre, saranno definiti opportuni indicatori in grado di misurare tali dimensioni (Paragrafo 3.1.2). Successivamente, saranno illustrati i passi operativi che un gestore di archivio amministrativo deve implementare, per poter misurare e valutare la qualità dei propri dati toponomastici (Paragrafo 3.1.3). Infine, il Paragrafo 3.1.4 sarà dedicato ad illustrare una metodologia statistica utile per identificare sottoinsiemi di indirizzi omogenei dal punto di vista dell'errore sintattico. Gli specifici sottoinsiemi che presentano un rischio elevato di errore sintattico saranno nel seguito indicati come *aree critiche* [3].

3.1.1 Aspetti Concettuali

Nella descrizione della misurazione della qualità dei dati toponomastici, si prenderà in esame l'insieme minimo di cinque componenti, ossia: provincia, comune, DUG, denominazione area di circolazione, numero civico, introdotte nel capitolo 2.

L'*accuratezza sintattica* è strettamente legata alla presenza o meno di errori nelle suddette componenti. Questa misura il grado di vicinanza tra il valore assunto dalle componenti di un indirizzo e i valori corretti o ufficiali delle componenti stesse.

La misurazione della vicinanza tra i valori *osservati* delle componenti di un indirizzo e i valori *corretti* delle stesse, implica l'esistenza di un vocabolario di *valori ufficiali*. L'esistenza di questo vocabolario è strettamente connessa agli aspetti normativi illustrati in dettaglio in Appendice 2. Esiste un vocabolario delle denominazioni ufficiali delle province e dei comuni italiani. Tale vocabolario, riportato sul sito http://www.istat.it, è gestito dall'ISTAT che, a tal fine, raccoglie le indicazioni fornite

[3] I concetti sviluppati nel capitolo 3 derivano da uno studio preliminare, svolto da un sottogruppo di lavoro (composto da: Marcello D'Orazio, Piero Demetrio Falorsi, Antonio Pavone, Marina Signore, Giorgia Simeoni), nell'ambito dell'accordo di collaborazione AIPA-ISTAT, riportato nell'appendice 3 del volume.

dagli Enti (Ministero degli Interni e Regioni) che hanno competenza sulle denominazioni in oggetto. Invece, i vocabolari con i valori ufficiali delle componenti DUG, denominazione dell'area di circolazione, e numero civico sono gestiti a livello locale dai singoli comuni, i quali sono anche gli enti responsabili di tali componenti. Attualmente, non esiste un vocabolario che integri i diversi vocabolari gestiti a livello locale. In mancanza di un vocabolario integrato, è disponibile lo *Stradario Nazionale*. Esso è realizzato dall'Istat in occasione dei Censimenti Generali della Popolazione, in base alle indicazioni fornite dai comuni. Attualmente, esiste la versione dello Stradario Nazionale relativa al Censimento del 1991. Nella sperimentazione condotta (descritta nel Paragrafo 3.2), si è utilizzata una versione aggiornata al 1999 dello Stradario Nazionale, realizzata dalla società SEAT Pagine Gialle. Nel corso dell'anno 2005 sarà messa a punto dall'ISTAT la nuova versione dello Stradario aggiornata con i dati del Censimento 2001.

Gli errori sintattici, che possono riguardare le componenti singolarmente considerate, riguardano la mancanza del valore o la presenza di un valore non ammissibile. La mancanza di un valore definisce un *errore di completezza*. La presenza di un valore non ammissibile si può verificare quando il valore osservato per una componente non coincide con quello ufficiale; ad esempio, si ha un errore dovuto ad un valore non ammissibile se per l'area di circolazione "*Giuseppe Garibaldi*", si registra il valore "*Giuseppe Gariboldi*".

Esaminando, invece, l'indirizzo come l'insieme delle cinque componenti sopra riportate, bisogna considerare che queste non sono tra loro indipendenti, ma sono in *relazione strettamente gerarchica*. Da ciò deriva che, l'accuratezza sintattica di un indirizzo non può prescindere dalla *consistenza interna* delle sue componenti. Ciò implica che la combinazione di due o più valori delle componenti è sintatticamente errata se tali valori sono tra loro inconsistenti, anche se presi singolarmente sono sintatticamente corretti. Un esempio è rappresentato da un indirizzo in cui il comune e la provincia hanno entrambi valori ammissibili, ma il comune non appartiene alla provincia indicata. Situazioni di questo tipo danno origine a *errori di consistenza interna*. I possibili casi di errore di consistenza interna sono essenzialmente i seguenti:

- Caso 1: provincia e comune inconsistenti; il valore del comune, pur essendo ammissibile, non è coerente con quello della provincia: cioè nell'ambito della provincia dell'indirizzo non esiste alcun comune con la denominazione riportata nell'indirizzo stesso;
- Caso 2: provincia e comune consistenti tra loro, ma inconsistenti con l'area di circolazione; l'area di circolazione può esistere ma non è presente nel comune indicato nell'indirizzo. Un caso particolare si verifica quando nel comune esiste un'area di circolazione con la medesima denominazione di quella riportata nell'indirizzo, ma la DUG indicata nell'indirizzo non è consistente con la denominazione. Si consideri ad esempio che in un determinato indirizzo sia riportata come area di circolazione la dizione "Piazza Giuseppe Garibaldi"; il caso di inconsistenza interna appena illustrato può avvenire se nel comune dell'indirizzo, l'area di circolazione di cui sopra non esiste, mentre esiste "Via Giuseppe Garibaldi";

– Caso 3: provincia, comune e area di circolazione consistenti tra loro, ma inconsistenti con il numero civico; nell'area di circolazione indicata nell'indirizzo non esiste un numero civico uguale a quello riportato nell'indirizzo.

Nella pratica, per valutare l'accuratezza sintattica dell'indirizzo nel suo complesso non si può prescindere dalla considerazione dello strumento che si utilizza per mettere in relazione gli indirizzi di un archivio con quelli ufficiali presi come riferimento. L'operazione in grado di stabilire una corrispondenza tra un dato indirizzo e il valore ufficiale di riferimento prende il nome di *riconoscimento* ed è realizzata tramite opportuni algoritmi [SISTER2004] incorporati in specifici software di *riconoscimento e normalizzazione degli indirizzi*. Tali software sono in genere articolati in due distinti passi: (*i*) nel primo passo, di *riconoscimento*, gli indirizzi dell'archivio sono confrontati con gli indirizzi riportati su uno stradario ufficiale, distinguendo tra due sottoinsiemi costituiti dall'insieme degli *indirizzi riconosciuti* e da quello degli *indirizzi non riconosciuti*; (*ii*) nel secondo passo, di *normalizzazione*, per gli *indirizzi riconosciuti* è generato un archivio in cui gli indirizzi originari sono affiancati da indirizzi corretti riportati in un formato standard; tali indirizzi sono indicati come *indirizzi normalizzati*. Gli algoritmi incorporati nei software sono in grado *di riconoscere* l'esatta dizione di un indirizzo anche per particolari casi di violazione dell'accuratezza sintattica. La qualità del riconoscimento, in tali casi, deve essere opportunamente esaminata. In effetti, la procedura automatizzata è soggetta al rischio, di produrre i cosiddetti *falsi riconoscimenti.*

Il concetto di riconoscimento di un indirizzo permette di qualificare in modo opportuno, le violazioni dell'accuratezza sintattica: a tal fine è utile definire una classificazione degli indirizzi che tenga conto contemporaneamente delle violazioni dell'accuratezza sintattica e dell'*esito* della procedura di riconoscimento; si distingue quindi tra:

– indirizzo corretto; l'indirizzo corrisponde esattamente al valore ufficiale e viene riconosciuto dal software;

– indirizzo che presenta una violazione debole dell'accuratezza sintattica; l'indirizzo non corrisponde esattamente al suo valore ufficiale, ma tale violazione dell'accuratezza non influisce negativamente sul riconoscimento dell'indirizzo da parte del software;

– indirizzo che presenta una violazione forte dell'accuratezza sintattica; l'indirizzo non corrisponde al suo valore ufficiale e tale violazione dell'accuratezza non permette il riconoscimento dell'indirizzo da parte del software.

A chiusura di questo Paragrafo è utile notare che la violazione dell'accuratezza sintattica di un indirizzo può dipendere anche da problemi di *aggiornamento*. In effetti i dati toponomastici ufficiali variano nel tempo: si creano nuovi comuni o province, le strade sono ridenominate, etc.. Per cui può essere che un indirizzo, sintatticamente corretto al momento della sua acquisizione nell'archivio amministrativo, non mantenga questa sua caratteristica con il trascorrere del tempo. Per tale ragione, al fine di migliorare la qualità dei dati toponomastici, è opportuno dedicare particolare attenzione alle azioni che consentono di aggiornare i dati stessi; la descrizione di uno schema generale utile a guidare tali azioni è descritto in dettaglio nel capitolo 5.

3.1.2 Indicatori di Accuratezza Sintattica

In questo Paragrafo saranno formalizzate le espressioni di alcuni indicatori atti a misurare l'accuratezza sintattica dei dati toponomastici negli archivi amministrativi. In particolare, il Paragrafo 3.1.2.1 è dedicato ad esplicitare i più importanti indicatori per misurare la violazione dell'accuratezza sintattica; mentre, nel Paragrafo 3.1.2.2, è dedicato a formalizzare indicatori in grado di misurare alcune specifiche dimensioni dell'accuratezza sintattica, ossia la completezza e la consistenza; infine, nel Paragrafo 3.1.2.3, si descrivono indicatori per specifici sottoinsiemi di record appartenenti ad un archivio amministrativo.

3.1.2.1 Indicatori Generali di Accuratezza Sintattica

Un modo per valutare l'errore sintattico di un dato toponomastico, gerarchicamente consistente, è quello di segnalare qualsiasi tipo di violazione rispetto a ciascuna componente costituente l'insieme minimo.

Un primo insieme di indicatori, utili a denotare la violazione dell'accuratezza sintattica per ciascuna delle componenti costituenti l'insieme minimo, è costituito dagli indici di *Errore Sintattico per Componente* (*ESC*). Al fine di illustrare la formula di calcolo di tali indici, si consideri una specifica base di dati pubblica che memorizza indirizzi e si indichi con: *N*, il numero di record[4] della base di dati contenenti indirizzi; *i*, il generico record (i=1,…,N); *j* (j=1,…,5), la specifica componente dell'indirizzo (facente parte dell'insieme minimo); $y_{i,j}$ una variabile indicatrice che assume valore 1 se la componente *j* del dato toponomastico del record *i* è sintatticamente non corretta, e valore 0 altrimenti.

Avendo definito tali quantità è possibile quindi pervenire alla definizione dell' indice di *Errore Sintattico per la Componente j* (*ESCj*): come:

$$ESCj = \frac{1}{N}\sum_{i=1}^{N} y_{i,j}\ , \qquad j=1,\ldots,5. \tag{3.1}$$

Mediante la (3.1) vengono quindi costruiti i cinque indicatori: *ESC1* (proporzione di errori nella provincia), *ESC2* (proporzione di errori nel comune), *ESC3* (proporzione di errori nella DUG), *ESC4* (proporzione di errori nella denominazione dell'area di circolazione) e *ESC5* (proporzione di errori nel numero civico).

Per valutare la relazione tra accuratezza sintattica di un indirizzo e il riconoscimento o meno dello stesso da parte dell'apposito software risulta conveniente introdurre la variabile binaria r_i (i=1,…,N), dove:

$$r_i = \begin{cases} 0, & \text{se l'}i\text{ - esimo indirizzo è stato riconosciuto} \\ 1, & \text{altrimenti} \end{cases}.$$

[4] Nel seguito, si farà riferimento al concetto di *record* in senso generale. In ambiente relazionale, i record sono assimilabili alle *tuple* (o righe) di tabelle relazionali.

L'utilizzo congiunto delle variabili $y_{i,j}$ e r_i permette di costruire tre *tassi* utili a descrivere in modo *generale* la relazione esistente tra l'esito della procedura di riconoscimento e l'accuratezza sintattica. Definiamo i seguenti indicatori:

- Indicatore Generale di Esito Corretto (IGEC)

$$IGEC = \frac{1}{N}\sum_{i=1}^{N}\delta\left((s_i = 0) \wedge (r_i = 0)\right) \qquad (3.2)$$

- Indicatore Generale di Esito di Violazione Debole (IGEVD) dell'accuratezza sintattica

$$IGVD = \frac{1}{N}\sum_{i=1}^{N}\delta\left((s_i > 0) \wedge (r_i = 0)\right) \qquad (3.3)$$

- Indicatore Generale di Esito di Violazione Forte (IGEVF) dell'accuratezza sintattica

$$IGVF = \frac{1}{N}\sum_{i=1}^{N}\delta\left((s_i > 0) \wedge (r_i = 1)\right), \qquad (3.4)$$

dove $s_i = \sum_{j=1}^{5} y_{i,j}$ e $\delta(\cdot)$ è una funzione binaria che assume valore 1 se la condizione in parentesi è verificata e valore 0 altrimenti.

L'indicatore *IGEC* descrive la frequenza relativa degli indirizzi corretti. L'indicatore *IGVD* denota la frequenza relativa di indirizzi con almeno un errore sintattico riconosciuti dal software; infine l'indicatore *IGVF* indica la frequenza relativa di indirizzi con almeno un errore sintattico che impedisce il riconoscimento da parte del software.

Per riuscire ad analizzare in modo più *specifico* la relazione esistente tra *accuratezza sintattica* e *esito della procedura di riconoscimento* è utile costruire per il singolo indirizzo i una nuova variabile, ϕ_i, mediante la concatenazione delle variabili indicatrici $y_{i,j}$ (j=1,…,5) e r_i. Questa variabile, di tipo vettoriale, è composta da sei posizioni. Le prime cinque posizioni, occupate dalle variabili $y_{i,j}$, assumono valore 1 o 0 a seconda che la componente, cui la posizione si riferisce (prima posizione a sinistra per la provincia, seconda posizione per il comune, terza posizione per la DUG, quarta posizione per la denominazione dell'area di circolazione, la quinta posizione per il numero civico), presenti o meno un valore sintatticamente non corretto. La sesta posizione è occupata dalla variabile r_i. Così, ad esempio, $\phi_i = 000000$ indica che il record i-esimo (i=1,…,N) presenta dei valori sintatticamente corretti per tutte le cinque componenti ed è stato riconosciuto dal software; al contrario, $\phi_i = 000101$ indica che l'indirizzo, che non è stato

riconosciuto dal software, presenta valori esatti per tutte le componenti ad eccezione di quella in posizione 4, ovvero la *denominazione dell'area di circolazione*. La variabile ϕ_i può presentare 2^6=64 possibili valori o *configurazioni*. Un'analisi specifica della relazione tra *errore sintattico* e *esito della procedure di riconoscimento* può essere effettuata contando il numero di occorrenze assolute e relative di ogni possibile configurazione. Per illustrare in modo formale la costruzione degli *Indici Specifici di Esito* (*ISE*), si ordinino lessicograficamente le configurazioni in parola e si denoti con ${}_c\phi$ (c =1,…,64) la configurazione che occupa la *c-esima* posizione nell'ordinamento lessicografico suddetto; ad esempio, $_1\phi = 000000$, $_2\phi = 000001, \ldots, {}_{64}\phi = 111111$. *L'Indice Specifico di Esito* per la generica configurazione ${}_c\phi$ ($ISE_{\,c}\phi$) viene quindi calcolato come:

$$ISE_{\,c}\phi = \frac{1}{N}\sum_{i=1}^{N}\delta\left(\phi_i = {}_c\phi\right). \quad (c=1,\ldots,64) \qquad (3.5)$$

Mediante la (3.5) vengono quindi costruiti i 64 indicatori: ISE000000 (proporzione di indirizzi di record completamente esatti e riconosciuti), …, ISE111111 (proporzione di indirizzi con le cinque componenti errate e non riconosciuti dal software). E' importante notare che, benché il numero teorico di indici ISE sia pari a 64, il numero effettivo sia molto minore; tale numero, come si nota dalla tabella 3.7 riportata nel paragrafo 3.2, è pari esattamente a 29. Ciò, come si può desumere da quanto illustrato nello stesso paragrafo 3.2, è dovuto alle problematiche connesse alla costruzione delle stringhe ϕ_i, a partire dagli output rilasciati dai software di riconoscimento e normalizzazione degli indirizzi.

Le misure precedenti si basano su un criterio stringente di identità di tipo vero o falso e non forniscono alcuna gradazione quando è accertata una violazione debole. Una misura più fine del grado di accuratezza sintattica di un dato toponomastico può essere ottenuta, rifacendosi a quanto riportato nella letteratura sviluppata nell'ambito del *record linkage* (vedi [Winkler2004]) relativamente agli algoritmi utili a confrontare stringhe testuali (vedi [Bilenko2003], [Cohen2003], [Navarro2001]). L'algoritmo, che è stato sviluppato per verificare il grado di vicinanza di un indirizzo al proprio valore normalizzato, si articola nei seguenti passi:

- Passo 1: si concatenano in un unico campo, dopo aver rimosso gli spazi vuoti, tutte le componenti dell'indirizzo da valutare e si effettua la medesima operazione per l'indirizzo normalizzato, ossia l'indirizzo ottenuto come output del software di riconoscimento e di normalizzazione. La rimozione degli spazi vuoti serve ad evitare eventuali problemi connessi ai differenti formati di acquisizione delle informazioni in fase di inserimento, che possono anche non prevedere la possibilità d'includere spazi;
- Passo 2: si definisce una lunghezza prefissata, ad esempio 3, e si individuano nella stringa (ottenuta al passo 1) dell'indirizzo normalizzato tutte le sottostringhe di tale lunghezza costituite da sequenze consecutive di caratteri. Tale operazione viene effettuata anche nella stringa relativa all'indirizzo da valutare;

- Passo 3: si pone pari a 0 uno specifico contatore. Si adotta un algoritmo di tipo sequenziale. Si considera la prima delle sottostringhe (identificate al passo 2) della stringa relativa all'indirizzo normalizzato e la si confronta in modo sequenziale con ciascuna delle sottostringhe dell'indirizzo da valutare. Ogni volta che si verifica l'uguaglianza delle sottostringhe si incrementa di uno un contatore. La sequenza viene iterata per la seconda sottostringa dell'indirizzo normalizzato, fino a considerare l'ultima sottostringa. Ad ogni iterazione viene aggiornato il valore del contatore;
- Passo 4: l'indicatore complessivo di accuratezza sintattica per il generico indirizzo i è ottenuto come rapporto tra il valore del contatore, calcolato al passo precedente, e il numero di sottostringhe dell'indirizzo normalizzato identificate al passo 2. Si noti che questa procedura può essere adottata solo nel caso in cui l'indirizzo è riconosciuto; in caso contrario infatti, l'indicatore complessivo di accuratezza sintattica è posto uguale a zero.

E' utile introdurre un esempio per meglio esplicitare i passi di cui sopra illustrati. Consideriamo l'indirizzo da valutare e il suo valore normalizzato illustrato in figura 3.1.

Indirizzo da valutare												
Componente	Valore della componente											
Provincia	T	V										
Comune	F	O	N	T	E	O	N	E		‘		-
DUG	V	I	A									
Denominazione della area di circolazione	S	A	N	N	I	C	O	L	O		‘	
Numero Civico	1											
Indirizzo normalizzato												
Componente	Valore della componente											
Provincia	T	V										
Comune	F	O	N	T	E							
DUG	V	I	A									
Denominazione della area di circolazione	O	N	E	‘								
Numero Civico	1											

Figura 3.1: Esempio di indirizzo

Esempio: Passo 1

Nel passo 1, separatamente per ciascun indirizzo, le componenti sono concatenate, eliminando gli spazi vuoti, formando le due stringhe riportate in figura 3.2.

Indirizzo da valutare

T	V	F	O	N	T	E	O	N	E	‘	-	V	I	A	S	A	N	N	I	C	O	L	O	‘	1

Indirizzo normalizzato

T	V	F	O	N	T	E	V	I	A	O	N	E	‘	1

Figura 3.2: Esempio di formazione di stringhe

Esempio: Passo 2

Definendo una lunghezza pari a tre, si individuano le seguenti 13 *sottostringhe* nella stringa relativa all’indirizzo normalizzato:

TVF, VFO, FON, ONT, NTE, TEV, EVI, VIA, IAO, AON, ONE, NE’, E‘1.

Si noti che il numero delle sottostringhe (13) è ottenuto mediante la formula

13 = 15 - 3+1

dove 15 è il numero totale di caratteri della stringa normalizzata e 3 è il numero di caratteri (lunghezza) delle sottostringhe.

Inoltre, applicando la stessa formula, si individuano le seguenti 23 sottostringhe nella stringa dell’indirizzo da valutare:

TVF, VFO, FON, ONT, NTE, TEO, EON , ONE, NE‘, E‘-, ‘-V, -VI, VIA, IAS, ASA, SAN, ANN, NIC, ICO, COL, OLO, LO‘, O‘1.

Esempio: Passo 3

Si pone pari a 0 il valore di uno specifico contatore. Si considera la prima sottostringa, TVF, dell’indirizzo normalizzato e si pone a confronto, in modo sequenziale con tutte le sottostringhe dell’indirizzo da valutare; si nota una relazione di uguaglianza solo con la prima sottostringa di tale insieme. Il valore del contatore è aggiornato con il valore corrispondente al numero di volte in cui si verifica la relazione di uguaglianza, ossia, nel caso in esame, pari a 1. Si considera quindi la seconda sottostringa dell’indirizzo da normalizzare, VFO e si pone a confronto, in modo sequenziale con tutte le sottostringhe dell’indirizzo da valutare. Anche in questo caso l’uguaglianza è verificata una sola volta. Il contatore è aggiornato aggiungendo tale occorrenza al valore precedente. Si prosegue in modo sequenziale, fino a considerare l’ultima sottostringa E‘1. Al termine di questa procedura, il contatore assumerà un valore pari a 8. Questo valore esprime il numero di volte che le sottostringhe dell’indirizzo normalizzato sono uguali alle sottostringhe dell’indirizzo da valutare.

Esempio: Passo 4

Per l'indirizzo preso in esame, l'*indicatore complessivo di accuratezza sintattica* è pari a 8/13.

Un elemento cruciale di tale procedura è quello della definizione della lunghezza delle sottostringhe. Dopo una serie di analisi empiriche, si è appurato che la lunghezza ottimale, nel caso di dati toponomastici, è pari a tre. Infatti, se si confronta la stringa normalizzata con quella da valutare un carattere per volta, si rischia di assegnare punteggi elevati di accuratezza, poiché è molto elevata la possibilità che ciascun elemento possa ripresentarsi più volte nell'indirizzo da valutare. Lo stesso vale se si considerano coppie consecutive di caratteri. Meno rischiosa è la scansione se si utilizzano sequenze consecutive di caratteri composte da triple adiacenti. Una lunghezza pari a 4, o oltre, è rischiosa in quanto tende a sottostimare l'accuratezza effettiva di un indirizzo. Ad esempio, se consideriamo la quadrupla di caratteri VIAO nell'indirizzo normalizzato, la scansione sull'indirizzo da valutare, darebbe esito negativo, in quanto esiste solo la quadrupla VIAS e quindi non si accerterebbe l'uguaglianza tra le DUG dei due indirizzi.

Al fine di esprimere in modo formale la procedura sopra illustrata, facendo riferimento all'indirizzo *i-esimo*, si indichi con: J_i ($J_i \geq 3$) il numero di caratteri presenti nella stringa dell' *indirizzo da valutare*; x_{ij} il valore assunto dal *j*-esimo (j=1,…, J_i) carattere della stringa dell'indirizzo da valutare; $\mathbf{x}_i = \left(x_{i1},...,x_{ij},...,x_{iJ_i}\right)'$ il vettore contenente i J_i caratteri x_{ij}; ${}_x\boldsymbol{\omega}_{ij} = \left(x_{ij}, x_{ij+1}, x_{ij+2}\right)'$ (con $j \leq J_i - 2$) un vettore contenente il carattere x_{ij} e i due caratteri ad esso consecutivi; K_i ($K_i \geq 3$) il numero di caratteri presenti nella stringa dell'*indirizzo normalizzato*; v_{ik} il valore assunto dal *k*-esimo (k=1,…, K_i) carattere della stringa dell'indirizzo normalizzato; $\mathbf{v}_i = \left(v_{i1},...,v_{ik},...,v_{iK_i}\right)'$ il vettore contenente i K_i caratteri v_{ik}; ${}_v\boldsymbol{\omega}_{ik} = \left(y_{ik}, y_{ik+1}, y_{ik+2}\right)'$ (con $j \leq J_i - 2$) il vettore stringa di tre caratteri contenente il carattere v_{ik} e i due caratteri ad esso consecutivi. Le stringhe $\mathbf{x}_i$ e $\mathbf{v}_i$ non contengono caratteri spazio a loro interno.

L'indice di accuratezza sintattica relativo all'indirizzo *i* può essere definito come:

$$AS_i = \frac{1}{K_i - 2}(1 - r_i) \sum_{j=1}^{J_i-2} \sum_{k=1}^{K_i-2} \delta\left({}_v\boldsymbol{\omega}_{ik} = {}_x\boldsymbol{\omega}_{ij}\right) \qquad (3.6)$$

in cui, si ricorda che, r_i denota la variabile dicotomica che assume valore 1 se l'indirizzo i non è stato riconosciuto dalla procedura automatizzata di normalizzazione. L'indice AS_i assume valore 1 nel caso in cui si verifica la doppia condizione: (*i*) $\mathbf{v}_i = \mathbf{x}_i$ e (*ii*) non sono presenti triple ${}_v\boldsymbol{\omega}_{ik}$ che assumono lo stesso valore più di una volta. In simboli tale condizione è data da ${}_v\boldsymbol{\omega}_{ik} \neq {}_v\boldsymbol{\omega}_{ik'}$ $\left(\text{per } k \neq k' \text{ e } (k,k') = 1,...K_i - 2\right)$.

La seconda condizione può essere compresa con un esempio. Si consideri la seguente denominazione dell'area di circolazione "*SANTABARBARA*". La tripla BAR si ripete due volte e dunque nell'indirizzo da valutare, se corretto, sarà conteggiato 4 volte. In tal caso il numero di occorrenze osservate come uguali sarà superiore al numero di triple possibili dell'indirizzo normalizzato. In tal caso il punteggio di accuratezza sarà superiore ad uno.

Nel caso in cui sia violata la condizione (*ii*), l'indice AS_i può assumere valori superiori ad 1. La presenza di triple ripetute determina una sopra valutazione dell'indice AS_i. Tuttavia, l'analisi empirica ha mostrato un numero limitato di tali casi. Per risolvere almeno parzialmente questo problema si propone una versione leggermente modificata della (3.6):

$$AS_i = \min\left(1, \left(\frac{1}{K_i - 2}(1 - r_i)\sum_{j=1}^{J_i-2}\sum_{k=1}^{K_i-2}\delta\left({}_v\omega_{ik} = {}_x\omega_{ij}\right)\right)\right). \qquad (3.7)$$

L'indice AS_i oltre ad essere una misura sintetica della qualità di un dato toponomastico, fornisce anche un segnale sulla qualità del riconoscimento effettuato dal software. Infatti, nel caso l'indice assuma valori esigui è opportuno verificare la correttezza del riconoscimento ottenuto dalla procedura di normalizzazione, in quanto potrebbe trattarsi di un *falso riconoscimento*.

L' *Indice Generale di Accuratezza Sintattica* (*AGAS*) di un archivio può essere quindi ottenuto come semplice media degli indici di accuratezza sintattica AS_i calcolati per le singole unità:

$$AGAS = \frac{1}{N}\sum_{i=1}^{N} AS_i . \qquad (3.8)$$

3.1.2.2 Indicatori Specifici di Completezza e Consistenza

Nel Paragrafo precedente si sono illustrati una serie di indicatori atti a descrivere la *accuratezza sintattica* di un indirizzo; si esaminano ora alcuni indicatori utili a chiarire specifici aspetti dell'accuratezza sintattica, ossia la *completezza* e la *consistenza*.

Si consideri, dapprima la completezza. Si riprenda la notazione introdotta nel Paragrafo precedente e si indichi con $z_{i,j}$ una variabile indicatrice che assume valore 1 se la componente *j* del dato toponomastico del record *i* è non completa, e valore 0 altrimenti. L'*Errore di Completezza per la Componente j* (*ECCj*) può essere ottenuto come la *frazione di valori mancanti* per la componente suddetta, ossia:

$$ECCj = \frac{1}{N}\sum_{i=1}^{N} z_{i,j} \qquad j=1,\ldots,5. \qquad (3.9)$$

Mediante la (3.9) sono ottenuti cinque indicatori: *ECC1* (proporzione di record che hanno valore mancante nella componente provincia), *ECC2* (proporzione di record

che hanno valore mancante nella componente comune), *ECC3* (proporzione di record che hanno valore mancante nella componente DUG), *ECC4* (proporzione di record che hanno valore mancante nella componente denominazione dell'area di circolazione), *ECC5* (proporzione di record che hanno valore mancante nella componente numero civico).

Il fatto che la completezza sia una specifica componente dell'accuratezza sintattica può essere verificato notando che la presenza di un errore di completezza sulla componente elementare (che si esplicita nella condizione $z_{i,j}=1$) implica, necessariamente, il verificarsi di un errore sintattico (che si sostanzia nella condizione $y_{i,j}=1$); viceversa la presenza di un errore sintattico, non implica un errore di completezza.

Un indicatore che fornisce una misura complessiva del grado di non completezza in una base di dati è l' *Errore di Completezza Medio* (*ECM*), definito nel seguente modo:

$$ECM = \frac{1}{N \times 5} \sum_{i=1}^{N} \sum_{j=1}^{5} z_{i,j} \,. \qquad (3.10)$$

Si esamini ora la *consistenza interna* e si riprenda quanto introdotto nel Paragrafo 3.1.1, si indichi con g_i una variabile indicatrice che assume valore 1 se per il record i si verifica almeno uno degli errori di consistenza individuati dai casi (1), (2) e (3) descritti nel Paragrafo suddetto, e valore 0 altrimenti. L'*Errore di Consistenza Interna* (*ECI*), può essere ottenuto come la frazione di record inconsistenti, ossia:

$$EGI = \frac{1}{N} \sum_{i=1}^{N} g_i \,. \qquad (3.11)$$

3.1.2.3 Indicatori di Accuratezza Sintattica per Specifici Sottoinsiemi di Record di un Archivio

Gli indicatori definiti nei due paragrafi precedenti fanno riferimento all'intero archivio. Tuttavia può essere interessante valutare gli indicatori stessi, non su tutti i record dell'archivio, ma per specifici sottoinsiemi di record dell'archivio stesso. Ad esempio, per un gestore di un archivio può essere rilevante conoscere se la qualità dei dati toponomastici dello specifico sottoinsieme dei record, in cui l'acquisizione degli indirizzi avviene tramite una procedura automatica, sia differente da quella degli altri record dell'archivio.

Dal punto di vista formale, indichiamo con ${}_dU$ uno specifico sottoinsieme di record dell'archivio, costituito da ${}_dN$ record, dove:

$${}_dN = \sum_{i=1}^{N} \delta\left(i \in {}_dU\right) \leq N \,. \qquad (3.12)$$

Gli indici di qualità riferiti allo specifico sottoinsieme ${}_dU$ sono nel seguito indicati denotando lo specifico indice di qualità con un pedice *d*, che individua il particolare

sottoinsieme di record dell'archivio cui è riferito l'indice. Gli indici di qualità per il sottoinsieme ${}_dU$ sono ottenuti mediante le formule descritte ai paragrafi 3.1.2.1 e 3.1.2.2 con la semplice sostituzione della quantità N con la quantità ${}_dU$ e la moltiplicazione di ciascuna variabile elementare riferita al singolo record i con la variabile indicatrice di appartenenza al sottoinsieme $\delta(i \in {}_dU)$; pertanto si ha:

$$ {}_d ESCj = \frac{1}{{}_d N} \sum_{i=1}^{N} y_{i,j} \ \delta(i \in {}_dU) \quad (j=1,\ldots,5) \qquad (3.13) $$

$$ {}_d IGEC = \frac{1}{{}_d N} \sum_{i=1}^{N} \delta\left((s_i = 0) \wedge (r_i = 0)\right) \delta(i \in {}_dU) \qquad (3.14) $$

$$ {}_d IGVD = \frac{1}{{}_d N} \sum_{i=1}^{N} \delta\left((s_i > 0) \wedge (r_i = 0)\right) \delta(i \in {}_dU) \qquad (3.15) $$

$$ {}_d IGVF = \frac{1}{{}_d N} \sum_{i=1}^{N} \delta\left((s_i > 0) \wedge (r_i = 1)\right) \delta(i \in {}_dU) \qquad (3.16) $$

$$ {}_d ISE_c\phi = \frac{1}{{}_d N} \sum_{i=1}^{N} \delta(\phi_i = {}_c\phi) \ \delta(i \in {}_dU) \quad (c=1,\ldots,64) \qquad (3.17) $$

$$ {}_d AGAS = \frac{1}{{}_d N} \sum_{i=1}^{N} AS_i \ \delta(i \in {}_dU) \qquad (3.18) $$

$$ {}_d ECCj = \frac{1}{{}_d N} \sum_{i=1}^{N} z_{i,j} \ \delta(i \in {}_dU) \qquad (3.19) $$

$$ {}_d ECM = \frac{1}{{}_d N \times 5} \sum_{i=1}^{N} \sum_{j=1}^{5} z_{i,j} \ \delta(i \in {}_dU) \qquad (3.20) $$

$$ {}_d EIG = \frac{1}{{}_d N} \sum_{i=1}^{N} g_i \ \delta(i \in {}_dU). \qquad (3.21) $$

Le precedenti espressioni costituiscono una forma generale; infatti, quando il sottoinsieme ${}_dU$ coincide con tutti i record dell'archivio, l'indice di qualità per la sottopopolazione corrisponde a quello dell'intero archivio, essendo ${}_d N = N$ e $\delta(i \in {}_dU) = 1$.

3.1.3 Passi Operativi per la Misurazione degli Indici di Qualità

Le formule computazionali degli indici di qualità sopra riportati prevedono da un punto di vista operativo i seguenti passi: *(i)* predisposizione e trattamento automatizzato dei dati toponomastici ai fini del riconoscimento e normalizzazione; *(ii)* trattamento degli output generati dal software di riconoscimento e normalizzazione

per computare gli indici descritti nel Paragrafo 3.1.2; *(iii)* analisi statistica di tali indicatori.

3.1.3.1 Predisposizione dei Dati Toponomastici

In generale, questa fase prevede una predisposizione della base di dati al fine di impiegare i software di normalizzazione. Ogni software presenta caratteristiche peculiari per la lettura dei dati, sebbene tutti siano in grado di leggere file ASCII in formato libero. Poiché questi software si basano sul confronto diretto di ciascun indirizzo con la base dei dati toponomastici ufficiali, l'operazione di riconoscimento e normalizzazione può comportare tempi lunghi di elaborazione. In tal caso, quando l'archivio amministrativo comprende un numero elevato di record, può essere conveniente estrarre un campione rappresentativo di tali record su cui valutare la qualità dei dati toponomastici. Infatti, le tecniche di campionamento offrono il vantaggio di ridurre i costi di produzione delle stime a fronte di una piccola perdita d'informazione. I successivi paragrafi saranno dedicati ad illustrare la predisposizione del campione e il trattamento dei dati con un software automatizzato di riconoscimento.

3.1.3.1.1 Selezione di un Campione Rappresentativo

Nel caso che un archivio nazionale contenga meno di 500.000 record è preferibile calcolare gli indici proposti nel Paragrafo 3.1.2 su tutti i record dell'archivio stesso. Viceversa, se l'archivio contiene almeno cinquecentomila record si suggerisce di estrarre un campione rappresentativo di tali record e di estendere, mediante opportune stime, le misure di qualità calcolate a tutto l'archivio.

L'utilizzo di un campione *casuale* degli indirizzi dell'archivio permette che: (*i*) i risultati della misurazione possano essere riferiti, in modo scientificamente fondato, a tutti gli indirizzi dell'archivio, anche alla porzione di quelli non inclusi nel campione; (*ii*) sia possibile calcolare la precisione campionaria delle stime degli indicatori di qualità calcolati sul campione.

Nel seguito si illustra brevemente la strategia di campionamento che potrebbe essere seguita. Vista la complessità del tema sul campionamento e delle possibili situazioni applicative che potrebbero essere incontrate, si daranno indicazioni con contenuto immediatamente operativo, valide solo per archivi di ampie dimensioni che contengano indirizzi relativi a tutto il territorio nazionale. Negli altri casi, il campione deve essere opportunamente calibrato e definito dalla collaborazione congiunta di esperti dell'archivio e di esperti del campionamento. Il lettore che voglia approfondire l'argomento può fare riferimento a uno dei testi specializzati, tra cui si segnala il seguente [Cicchitelli,Herzel92].

Il disegno di campionamento che si suggerisce di utilizzare per archivi di ampie dimensioni che contengano indirizzi relativi a tutto il territorio nazionale, è *stratificato con selezione delle unità di campionamento* (gli indirizzi) *negli strati senza reimmissione e a probabilità uguali*.

I passi operativi da seguire per la realizzazione del campione sopra illustrato sono di seguito elencati:

- Passo 1: definizione della stratificazione e suddivisione delle unità della popolazione in strati;
- Passo 2: definizione del numero totale di record, n, da includere nel campione e scelta del numero di record campione da assegnare al generico strato;
- Passo 3: selezione casuale del campione;
- Passo 4: calcolo delle stime degli indicatori di qualità;
- Passo 5: valutazione della precisione campionaria delle stime definite al punto precedente.

Per i primi tre passi, si fornisce un software di supporto in appendice.

Passo (1): Definizione della stratificazione e suddivisione delle unità della popolazione in strati

Dal punto di vista matematico, gli strati costituiscono una *partizione dei record dell'archivio di interesse*, nel senso che ciascun *record deve* appartenere a uno ed a un solo *strato.*

Gli strati possono essere definiti dalle modalità assunte da una variabile di tipo *geografico,* presente su tutti i record dell'archivio. Nel caso di grandi archivi nazionali si suggerisce di adottare come variabile di *stratificazione la provincia*; altre scelte di stratificazione possono essere costituite dalla *regione* o da particolari aggregazioni di comuni.

Nel seguito indichiamo con h il generico strato (h=1,…,H) e con N_h il numero di record dell'archivio appartenenti allo strato, essendo

$$\sum_{h=1}^{H} N_h = N \ .$$

Passo (2): Definizione del numero totale di record da includere nel campione e scelta del numero di record campione da assegnare al generico strato

Nella sperimentazione illustrata nel Paragrafo 3.2 è empiricamente emerso che una numerosità campionaria di 200.000 record è sufficientemente consistente ed è in grado di fornire stime affidabili degli indici di qualità.

Per semplificare le successive fasi di analisi statistica dei dati si suggerisce di definire il numero di record da selezionare nel generico strato h, in misura proporzionale alla numerosità della popolazione dello strato. In simboli, il numero n_h di record da selezionare nel generico strato h deve essere ottenuto come:

$$n_h = min\left(2, int\left(n\,\frac{N_h}{N}\right)\right) \qquad (h=1,\ldots,H) \qquad (3.22)$$

dove $int(\cdot)$ denota una funzione che arrotonda al valore intero più vicino, quanto racchiuso in parentesi. La formula precedente deve essere interpretata nel modo seguente: la numerosità del campione per strato è pari al valore arrotondato

dell'espressione $n\frac{N_h}{N}$; qualora tale valore sia inferiore ad 2, esso viene posto comunque pari a 2. Quest'ultima condizione fa emergere la necessità che gli strati siano definiti in modo che la loro numerosità N_h sia superiore a 2.

Così ad esempio, si supponga che un determinato archivio sia costituito da un numero di record N pari a 1.200.000. Si supponga inoltre di avere definito una numerosità globale del campione, n, pari a 200.000 record. Si consideri un generico strato costituito da 10.000 record. Il numero di record da selezionare nel campione dello strato suddetto è pari a 1667, come si evince dai seguenti passaggi:

$$n_h = min\left(2, int\left(200.000\,\frac{10.000}{1.200.0000}\right)\right) =$$

$$= min\left(2, int\left(1.666{,}666667\right)\right) =$$

$$= min\left(2, 1.667\right) = 1.667\,.$$

Passo (3): Selezione del campione

Gli n record appartenenti al campione devono essere selezionati mediante una procedura articolata nel modo seguente.

A ciascuna delle N unità dell'archivio deve essere attribuito un *numero casuale* generato da una distribuzione *casuale uniforme* nell'intervallo [0,1].

Gli N record dell'archivio sono quindi ordinati in modo crescente per *codice strato* e numero casuale.

Il campione del generico strato h (h=1,…,H), è costituito dai record che occupano le prime n_h posizioni nell'ordinamento dei record dello strato.

Passo (4): Calcolo delle stime degli indicatori di qualità

Le stime campionarie vengono ottenute nel modo di seguito descritto.

A ciascun record selezionato nel campione viene attribuito un *peso*, w_i , ottenuto come rapporto tra il numero di record dello strato, N_h, da cui è stato selezionato il record in oggetto e il numero di record selezionati nel campione, n_h, in formule

$$w_i = \frac{N_h}{n_h} \qquad \text{per} \quad i \in s_h \qquad\qquad (3.23)$$

dove s_h denota l'insieme dei record selezionati nello strato h. Dal punto di vista statistico, il peso in oggetto sta ad indicare che il record campione *rappresenta* se stesso e altri $(w_i - 1)$ record non osservati nel campione.

Riprendendo l'esempio introdotto al passo (2), a tutti i 1.667 record selezionati viene attribuito un peso w_i pari a $w_i = \frac{10.000}{1.667} = 5{,}9988$. Tale peso sta ad indicare che

ciascuno dei record incluso nel campione dello strato rappresenta se stesso ed altri 4,9988 record non inclusi nel campione dello strato.

Le stime degli indicatori di qualità vengono ottenute ponderando le variabili elementari, definite su ogni singolo record campionario, con il peso attribuito al record medesimo. Si descrivono nel seguito le formule delle stime degli indicatori, facendo riferimento al caso generale di stime per sottopopolazioni di record, illustrato nel Paragrafo 3.1.2.3:

$$ {}_d\tilde{E}SCj = \frac{1}{{}_d\tilde{N}} \sum_{h=1}^{H} \sum_{i \in s_h} w_i \, y_{i,j} \, \delta(i \in {}_dU) = $$

$$ = \frac{1}{{}_d\tilde{N}} \sum_{h=1}^{H} \sum_{i \in s_h} \frac{N_h}{n_h} y_{i,j} \, \delta(i \in {}_dU) \qquad (j=1,\ldots,5) \qquad (3.24) $$

$$ {}_d\tilde{I}GEC = \frac{1}{{}_d\tilde{N}} \sum_{h=1}^{H} \sum_{i \in s_h} \frac{N_h}{n_h} \, \delta\left((s_i = 0) \wedge (r_i = 0)\right) \delta(i \in {}_dU) \qquad (3.25) $$

$$ {}_d\tilde{I}GVD = \frac{1}{{}_d\tilde{N}} \sum_{h=1}^{H} \sum_{i \in s_h} \frac{N_h}{n_h} \, \delta\left((s_i > 0) \wedge (r_i = 0)\right) \delta(i \in {}_dU) \qquad (3.26) $$

$$ \tilde{I}GVF = \frac{1}{{}_d\tilde{N}} \sum_{h=1}^{H} \sum_{i \in s_h} \frac{N_h}{n_h} \, \delta\left((s_i > 0) \wedge (r_i = 1)\right) \delta(i \in {}_dU), \qquad (3.27) $$

$$ {}_d\tilde{I}SE_c\phi = \frac{1}{{}_dN} \sum_{h=1}^{H} \sum_{i \in s_h} \frac{N_h}{n_h} \, \delta\left(\phi_i = {}_c\phi\right) \delta(i \in {}_dU) \qquad (c=1,\ldots,64) \qquad (3.28) $$

$$ {}_d\tilde{A}GAS = \frac{1}{{}_d\tilde{N}} \sum_{h=1}^{H} \sum_{i \in s_h} \frac{N_h}{n_h} AS_i \, \delta(i \in {}_dU) \qquad (3.29) $$

$$ {}_d\tilde{E}CCj = \frac{1}{{}_d\tilde{N}} \sum_{h=1}^{H} \sum_{i \in s_h} \frac{N_h}{n_h} \, z_{i,j} \, \delta(i \in {}_dU) \qquad (3.30) $$

$$ {}_d\tilde{E}CM = \frac{1}{{}_d\tilde{N} \times 5} \sum_{h=1}^{H} \sum_{i \in s_h} \frac{N_h}{n_h} \sum_{j=1}^{5} z_{i,j} \, \delta(i \in {}_dU) \qquad (3.31) $$

$$ {}_d\tilde{E}IG = \frac{1}{{}_d\tilde{N}} \sum_{h=1}^{H} \sum_{i \in s_h} \frac{N_h}{n_h} \, g_i \, \delta(i \in {}_dU), \qquad (3.32) $$

dove ${}_d\tilde{N}$ denota la stima di ${}_dN$, ottenuta come

$$ {}_d\tilde{N} = \sum_{h=1}^{H} \sum_{i \in s_h} w_i \ \delta\left(i \in {}_dU\right) = \sum_{h=1}^{H} \sum_{i \in s_h} \frac{N_h}{n_h} \ \delta\left(i \in {}_dU\right). \qquad (3.33) $$

Passo (5): Calcolo della precisione delle stime

Ciascuna delle formule, dalla (3.24) alla (3.32), rappresenta la stima campionaria rispettivamente di ciascuna delle espressioni dalla (3.13) alla (3.21), ognuna delle quali costituisce il valore che sarebbe stato calcolato per uno specifico indicatore di qualità, qualora fosse stato possibile osservare tutti i record di un archivio amministrativo.

Le stime campionarie sono soggette a incertezza dovuta al fatto che si osserva solo un sottoinsieme dei record di interesse. Tale incertezza può essere adeguatamente misurata direttamente sui dati campionari. Per illustrare tali aspetti, si indichi con θ il valore incognito di uno degli indicatori di qualità introdotti nelle relazioni dalla (3.13) alla (3.21) e si denoti con $\tilde{\theta}$ la corrispondente stima campionaria. Una volta osservati i dati campionari è possibile calcolare la varianza campionaria degli stessi $\tilde{V}ar\left(\tilde{\theta}\right)$ mediante la relazione:

$$ \tilde{V}ar\left(\tilde{\theta}\right) = \sum_{h=1}^{H} \frac{N_h \left(N_h - n_h\right)}{n_h - 1} \sum_{i \in s_h} \left(\theta_i - \frac{1}{n_h} \sum_{i \in s_h} \theta_i \right)^2 \qquad (3.34) $$

dove θ_i indica una variabile linearizzata [Wolter1975], relativa al record *i-esimo* e che, come di seguito indicato, assume valore differente a seconda dell'indicatore di qualità considerato.

Avendo calcolato la stima $\tilde{\theta}$ e la corrispondente varianza $\tilde{V}ar\left(\tilde{\theta}\right)$, è possibile determinare gli estremi dell'intervallo di confidenza che con probabilità pari a 0,95 contengono il valore incognito θ:

$$ \Pr\left(\tilde{\theta} - 2\sqrt{\tilde{V}ar\left(\tilde{\theta}\right)} \le \theta \le \tilde{\theta} + 2\sqrt{\tilde{V}ar\left(\tilde{\theta}\right)}\right) = 0{,}95\,. \qquad (3.35) $$

Ad esempio, si supponga che la stima ${}_d\tilde{A}GAS$ sia pari a 0,87 e che la stima della varianza di tale indicatore $\tilde{V}ar\left({}_d\tilde{A}GAS\right)$ sia pari a 0,01. Mediante la (3.35) si può affermare che il valore vero incognito dell'indice ${}_dAGAS$, con probabilità pari a 0,95, sia compreso in un intervallo i cui estremi, inferiore e superiore, sono definiti rispettivamente da $0{,}87 - 2\sqrt{0{,}001} = 0{,}807$ (limite inferiore) e $0{,}87 + 2\sqrt{0{,}001} = 0{,}933$ (limite superiore).

A conclusione del Paragrafo, riportiamo le formule esplicite delle variabili elementari θ_i, a seconda degli indicatori di qualità considerati:

$$ \theta_i = \frac{1}{{}_d\tilde{N}} \left(y_{i,j} - {}_d\tilde{E}SCj \ \delta\left(i \in {}_dU\right) \right) \ \text{ per } \tilde{\theta} = {}_d\tilde{E}SCj $$

$$\theta_i = \frac{1}{{}_d\tilde{N}}\left(\delta\left((s_i = 0) \wedge (r_i = 0)\right) - {}_d\tilde{I}GEC\ \delta\left(i \in {}_dU\right)\right) \text{ per } \tilde{\theta} = {}_d\tilde{I}GEC$$

$$\theta_i = \frac{1}{{}_d\tilde{N}}\left(\delta\left((s_i > 0) \wedge (r_i = 0)\right) - {}_d\tilde{I}GVD\ \delta\left(i \in {}_dU\right)\right) \text{ per } \tilde{\theta} = {}_d\tilde{I}GVD$$

$$\theta_i = \frac{1}{{}_d\tilde{N}}\left(\delta\left((s_i > 0) \wedge (r_i = 1)\right) - {}_d\tilde{I}GVF\ \delta\left(i \in {}_dU\right)\right) \text{ per } \tilde{\theta} = {}_d\tilde{I}GVF$$

$$\theta_i = \frac{1}{{}_d\tilde{N}}\left(\delta\left(\phi_i = {}_c\phi\right) - {}_d\tilde{I}SE_c\phi\ \ \delta\left(i \in {}_dU\right)\right) \text{ per } \tilde{\theta} = {}_d\tilde{I}SE_c\phi$$

$$\theta_i = \frac{1}{{}_d\tilde{N}}\left(AS_i\ -\ {}_d\tilde{A}GAS\ \ \delta\left(i \in {}_dU\right)\right) \text{ per } \tilde{\theta} = {}_d\tilde{A}GAS$$

$$\theta_i = \frac{1}{{}_d\tilde{N}}\left(z_{i,j} -\ {}_d\tilde{E}CCj\ \delta\left(i \in {}_dU\right)\right) \text{ per } \tilde{\theta} = {}_d\tilde{E}CCj$$

$$\theta_i = \frac{1}{{}_d\tilde{N}5}\left(\sum_{j=1}^{5} z_{i,j} -\ {}_d\tilde{E}CM\ \delta\left(i \in {}_dU\right)\right) \text{ per } \tilde{\theta} = {}_d\tilde{E}CM$$

$$\theta_i = \frac{1}{{}_d\tilde{N}}\left(g_i -\ {}_d\tilde{E}IG\ \delta\left(i \in {}_dU\right)\right) \text{ per } \tilde{\theta} = {}_d\tilde{E}IG\,.$$

3.1.3.1.2 Trattamento dei dati con la procedura automatizzata di riconoscimento e normalizzazione

Si descrivono nel seguito le principali caratteristiche delle procedure informatiche di *riconoscimento e normalizzazione* degli indirizzi, indicate per brevità nel seguito come *software*. Tali software si compongono di due passi: (*i*) un passo di *riconoscimento* e (*ii*) un passo di *normalizzazione*.

Il passo di *riconoscimento* confronta l'indirizzo da valutare con *tutti* gli indirizzi presenti in una *base di dati* di indirizzi assunti come corretti o di riferimento. Per ciascun confronto possibile, si determina il valore di una funzione di distanza tra l'indirizzo da valutare e l'indirizzo corretto.

Nel caso che la funzione di distanza assuma valore pari a zero, allora significa che l'indirizzo da valutare è *esattamente uguale* (carattere per carattere) a uno specifico indirizzo di riferimento. In tal caso, si dice che l'indirizzo da valutare è riconosciuto per *uguaglianza.* In generale, nella banca dei dati ufficiali possono anche essere presenti vocabolari contenenti dizioni abbreviate, e riconosciute come corrette, di alcune specifiche componenti. In tali situazioni, il riconoscimento per uguaglianza può avvenire anche nel caso in cui la componente dell'indirizzo da valutare sia uguale ad una delle forme abbreviate presenti nei vocabolari di riferimento.

Nel caso che la funzione di distanza sia maggiore di zero, il *riconoscimento* può avvenire per *similitudine*. In tal caso, tra tutti gli indirizzi di riferimento, candidati al riconoscimento, è scelto quello che presenta la minima distanza rispetto all'indirizzo

da valutare. Affinché l'indirizzo da valutare sia riconosciuto per similitudine la distanza minima deve essere inferiore ad una determinata soglia di rischio. Più si aumenta il valore di questa soglia, maggiore è il rischio di ottenere falsi riconoscimenti, ossia che all'indirizzo da valutare sia erroneamente associato un indirizzo di riferimento. Ai fini del riconoscimento per similitudine, alcuni software utilizzano anche vocabolari contenenti sinonimi dei valori di riferimento di alcune specifiche componenti (ad esempio, località o aree di circolazione di uso comune, ma non riconosciute a livello ufficiale).

Nel passo di *normalizzazione*, a ciascun indirizzo riconosciuto è affiancato l'indirizzo di riferimento secondo il formato standard previsto dal software.

Una volta predisposti i record di input, secondo i formati di lettura previsti dal software, sarà necessario elaborarli per ottenere gli output attesi. Spesso ciò può comportare una pre-operazione di conversione dei dati in un formato consono alla sua lettura. In generale, tutti i software sono in grado di leggere file in formato ASCII fisso o libero.

Una volta ottenuti gli output del software è necessario importare tutti i dati (gli indirizzi da valutare e gli output generati dal software) in un apposito programma in grado di restituire le elaborazioni statistiche richieste. Programmi statistici indicati a tale scopo sono, ad esempio, SPSS (*Statistical Package for Social Science*) [SPSS2004] e SAS [SAS2004].

Una volta importati i dati nell'ambiente statistico prescelto, è necessario predisporre alcune routine di programma preliminari alla costruzione delle variabili indicatrici $y_{i,j}$ e r_i. La variabile r_i è fornita direttamente dal software. Quando $r_i = 1$, la variabile stringa ϕ_i può assumere i seguenti valori: (*i*) $\phi_i = 111111$ (nessuna componente dell'indirizzo è riconosciuta), (*ii*) $\phi_i = 011111$ (è riconosciuta solo la componente provincia), (*iii*) $\phi_i = 001111$ (sono riconosciute solo la componenti provincia e comune). In genere il software fornisce informazioni sufficienti in merito ai motivi di non riconoscimento di un indirizzo, che permettono di discriminare a quale delle tre situazioni sopra descritte appartiene l'indirizzo non riconosciuto. Una volta ricostruito il valore della variabile stringa ϕ_i è risolto il problema dei valori che assumono le singoli componenti $y_{i,j}$ che la costituiscono.

E' più complessa la problematica della attribuzione di un valore alle variabili $y_{i,j}$ quando $r_i = 0$. Dal punto di vista strettamente teorico, si attribuisce un valore pari a 0 alla variabile $y_{i,j}$ solo nel caso in cui la componente *j* dell'indirizzo da valutare sia identica alla stessa componente dell'indirizzo di riferimento.

Tuttavia, l'implementazione di questo criterio non è così semplice, come potrebbe apparire a prima vista, a causa dei diversi formati d'acquisizione dei dati toponomastici disponibili nei diversi archivi, così come per i formati di rilascio dei dati normalizzati. I problemi principali che un analista dovrà affrontare sono:

il software può prevedere un doppio formato, esteso ed abbreviato, per alcune componenti dell'indirizzo normalizzato;

per la forma abbreviata della componente DUG e della componente denominazione dell'area di circolazione, il software può prevedere il rilascio del dato in un unico campo;

il formato di acquisizione dell'archivio amministrativo non prevede la separazione, in campi distinti, di tutte o alcune componenti l'indirizzo.

Problema 1

In generale, i software di normalizzazione prevedono una doppia dicitura, estesa ed abbreviata, per le componenti comune, DUG e denominazione dell'area di circolazione.

A livello della *j-esima* componente, l'accertamento dell'errore sintattico deve essere effettuato sia rispetto alla forma estesa che rispetto alla forma abbreviata. Nel caso in cui una delle due forme corrisponda a quella riportata nell'indirizzo valutato, allora la variabile $y_{i,j} = 0$, altrimenti $y_{i,j} = 1$.

Per quanto riguarda l'indicatore sintetico di accuratezza AS_i possono essere ottenute 8 stringhe, egualmente corrette, per un indirizzo di riferimento. Nello specifico le 8 stringhe possibili sono riportate in Tabella 3.1.

Forme possibili	Provincia	Comune	DUG	Denominazione	Numero Civico
1	forma estesa	forma estesa	forma estesa	forma estesa	forma estesa
2	forma estesa	forma estesa	forma estesa	forma abbreviata	forma estesa
3	forma estesa	forma estesa	forma abbreviata	forma estesa	forma estesa
4	forma estesa	forma estesa	forma abbreviata	forma abbreviata	forma estesa
5	forma estesa	forma abbreviata	forma estesa	forma estesa	forma estesa
6	forma estesa	forma abbreviata	forma estesa	forma abbreviata	forma estesa
7	forma estesa	forma abbreviata	forma abbreviata	forma estesa	forma estesa
8	forma estesa	forma abbreviata	forma abbreviata	forma abbreviata	forma estesa

Tabella 3.1: Possibili forme di stringhe corrette per un indirizzo di riferimento

Ciascun indirizzo da valutare dovrà essere confrontato con tutte le 8 forme possibili dell'indirizzo di riferimento, producendo 8 indici AS_i. All'indirizzo *i-esimo* si

attribuirà il punteggio massimo tra gli 8 indici AS_i, poiché il massimo si riferisce alla stringa che meglio si adatta all'indirizzo da valutare.

In Figura 3.3 si riporta, a titolo di esempio, una routine sviluppata in SPSS, per l'ottenimento del punteggio di accuratezza sintattica AS_i.

```
* Determina l'ampiezza della stringa indirizzo da valutare meno 2.
COMPUTE temp_i = LENGTH(rtrim(ind_inp))-2.
* Ciclo per gli otto possibili indirizzi in output.
do repeat
        car_out=  temp_o1 to temp_o8 /
        str_out= ind_out1 ind_out 2 ind_out3 ind_out4 ind_out5 ind_out6 ind_out7 ind_out8 /
        car_corr = conta1 to conta8/
        indx = index1 to index8.
        * Inizializzazione del contatore.
        compute car_corr = 0.
        * Determina l'ampiezza della stringa indirizzo in output meno 2.
        COMPUTE  car_out= LENGTH(rtrim(str_out))-2 .
        loop x = 1 to car_out.
                loop y = 1 to temp_i .
                        * Se il software di normalizzazione riconosce l'indirizzo (r=0) allora conta le occorrenze
                         delle triple di caratteri.
                        if r = 0 and (SUBSTR(str_out,x,3) = SUBSTR(ind_inp,y,3))
                                car_corr = car_corr + 1.
                end loop.
        end loop .
        *Calcola l'indice.
        COMPUTE INDX = car_corr/car_out.
end repeat.
execute.
* individua il valore massimo tra gli otto possibili indici di accuratezza.
compute AS = max(index1, index2, index3, index4,  index5, index6, index7, index8).
execute.
* Se l'indice è superiore ad uno,  si pone AS uguale a uno.
if  AS> 1 AS=1.
execute.
```

Figura 3.3: Routine in SPSS per la determinazione del punteggio AS_i

Problema 2

Se il software restituisce in un unico campo la forma abbreviata dell'area di circolazione, si pone una complicazione per una corretta valutazione dell'accuratezza sintattica. Infatti, riprendendo il caso illustrato per il problema (1), se la forma estesa di un'area di circolazione è "Piazza Giuseppe Mazzini" e la sua forma abbreviata "P.zza G. Mazzini", allora sono da considerare corrette le due forme ibride "P.zza Giuseppe Mazzini" e "Piazza G. Mazzini". In tal caso, è necessario trovare il modo di separare il campo stringa unificato nelle sue due componenti primitive.

In linea di principio, se la DUG estesa possiede un complemento, anche la DUG abbreviata dovrebbe possederlo; a partire dall'informazione contenuta dalla DUG estesa e utilizzando come separatore lo spazio è possibile separare l'area di circolazione in DUG abbreviata e denominazione abbreviata.

Nei fatti, questo non è sempre verificato, poiché è possibile che il complemento sia presente nella DUG estesa e non in quella abbreviata. In tal caso, è necessario avvalersi di un algoritmo più complesso che, attraverso una serie di confronti con le forme estese e scomposizioni della forma abbreviata, è in grado di ottenere il risultato desiderato.

Di seguito, si propone una procedura scritta in SPSS che risolve questo problema.

```
*Verifica se esiste un complemento alla DUG estesa.
COMPUTE DUG_PAR = 0.
EXECUTE.
IF INDEX(RTRIM(LTRIM(DUG))," ") > 0 DUG_PAR= 1.
EXECUTE .
* Routine per la creazione dei campi DUG_A (DUG abbreviata) e DAC_A (denominazione area di circolazione abbreviata).
* I suffissi tmp, tm2 o t, indicano che la variabile assume un valore temporaneo di comodo.
        STRING DUGA_TMP(A10).
        COMPUTE DUGA_TMP = SUBSTR(LTRIM(VIA_UF_A),1,INDEX(LTRIM(VIA_UF_A)," ")).
        EXECUTE.
        STRING DACAB_T (A30).
        COMPUTE DACAB_T = SUBSTR(VIA_UF_A,LENGTH(RTRIM(DUGA_TMP))+2).
        EXECUTE.
        * Confronti tra stringhe.
        STRING DUGA_TM2 (A10).
        COMPUTE DUGA_TM2 = SUBSTR(DACAB_T,1, INDEX(LTRIM(DACAB_T)," ")).
        EXECUTE.
        * Crea DUG abbreviata.
        STRING DUG_A(A25).
        COMPUTE DUG_A = DUGA_TMP.
        IF (DUG_PAR = 1 AND
                (SUBSTR(LTRIM(RTRIM(DUGA_TM2)),1, LENGTH(LTRIM(RTRIM(DUG A_TM2))) ) <>
                SUBSTR(LTRIM(RTRIM(DAC_EST)),1,LENGTH(LTRIM(RTRIM(DUGA_TM2))))))
                DUG_A =CONCAT(LTRIM(RTRIM(DUGA_TMP))," ",LTRIM(RTRIM(DUGA_TM2))).
        EXECUTE.
        * Crea DAC abbreviata.
        STRING DAC_A (A25).
        COMPUTE DAC_A= SUBSTR(RTRIM(LTRIM(VIA_UF_A)),LENGTH(RTRIM(DUGA))+2).
        EXECUTE.
```

Figura 3.4: Routine per la divisione dell'area di circolazione in DUG abbreviata e Denominazione abbreviata

Problema 3

Molto più intricato è invece risolvere il problema della separazione in campi distinti quando l'indirizzo da valutare riporta in un unico campo il *punto sul territorio* (rappresentato dall'unione delle componenti DUG, denominazione area di circolazione e numero civico).

Una strategia, per risolvere in modo computazionalmente semplice il problema di valorizzare le variabili $y_{i,j}$ è di seguito descritta. Si considera la stringa relativa alla componente rilasciata dalla procedura di normalizzazione e si esamina se essa sia inclusa come sottostringa della stringa relativa al punto sul territorio dell'indirizzo da valutare. Questa operazione è eseguita sia per la forma estesa sia per la forma abbreviata delle componenti. Nel caso che una delle due forme sia inclusa come sottostringa della stringa relativa al punto sul territorio dell'indirizzo da valutare, la variabile $y_{i,j}$ corrispondente alla componente esaminata assume valore 0, altrimenti assume valore 1.

Tuttavia, questo criterio di verifica della violazione dell'errore sintattico, non basandosi su un criterio stringente d'identità, può non rilevare alcune situazioni di errore.

Si riprenda l'esempio introdotto nel Paragrafo 3.1.2.1 e si supponga che l'indirizzo da valutare e quello normalizzato siano riportati secondo i formati illustrati in figura 3.5.

Si prenda la denominazione area di circolazione "ONE'". La procedura di accertamento dell'errore sintattico proposta, verifica che tale valore sia presente, come sottostringa nel campo *punto sul territorio* fornito in input. Poiché questa condizione è vera, indipendentemente dagli altri valori presenti nel campo, si accerta l'assenza di errore sintattico nella denominazione area di circolazione. L'esempio riportato è importante per interpretare gli indicatori proposti in modo corretto. Questi

rilasciano solo parti di informazioni e l'accertamento di una violazione dell'errore sintattico implica che la lettura sia contestuale. Così, nell'esempio, è ovvio che tutte le componenti normalizzate, costituenti l'indirizzo da valutare, sono presenti, infatti $\phi_i = 000000$. Tuttavia, l'indice di accuratezza sintattica sull'indirizzo da valutare indica che, o la sequenza formale dell'indirizzo da valutare non è corretta, oppure sono presenti simboli lessicografici di localizzazione, non essenziali $AS_i = (8/13) < 1$.

Indirizzo da valutare	
Componente	**Valore della componente**
Provincia	TV
Comune	FONTE
Punto sul territorio	ONE' - VIA SAN NICOLO' 1
Indirizzo normalizzato	
Provincia	TV
Comune	FONTE
DUG	VIA
Denominazione area di circolazione	ONE'
Civico	1

Figura 3.5: Esempio di problema (3)

Quando l'indirizzo da valutare riporta in un unico campo il punto sul territorio, non è possibile verificare, in modo automatico, violazioni alla dimensione della completezza, nelle componenti che lo costituiscono. In particolare, non è possibile discriminare tra una componente mancante e una componente errata quando si ricerca una singola componente come sottostringa dell'indirizzo da valutare.

3.1.4 L'Analisi della Accuratezza Sintattica mediante un Modello di Segmentazione Regressiva

Dato un archivio amministrativo per cui sia stata valutata l'accuratezza sintattica, è opportuno intraprendere azioni di correzione di tali errori. Una strategia che si propone di correggere gli errori, analizzando i record singolarmente, sarebbe particolarmente onerosa. Pertanto, si propone di utilizzare una tecnica che consente di evidenziare delle aree dell'archivio caratterizzate da un'elevata concentrazione dell'errore sintattico. Per identificare le aree a più elevato rischio di errore sintattico si propone una tecnica statistica non parametrica, conosciuta come *alberi di regressione* [Breiman84].

La tecnica è applicabile qualunque sia la natura delle variabili utilizzate (ordinale cardinale, continua, ecc..). Queste sono suddivise in due tipologie:

– la variabile dipendente, ossia la variabile d'interesse che si vuole interpretare ;

– le variabili esplicative o predittive (covariate) mediante le quali si cerca di individuare un modello esplicativo del comportamento della variabile dipendente.

Il metodo ripartisce un collettivo (nel caso in esame i record del campione o dell'intero archivio amministrativo) in gruppi (o nodi) sempre più omogenei al loro interno, rispetto alla variabile dipendente, mediante una progressiva divisione dicotomica delle variabili esplicative**.**

Al primo passo, l'intero collettivo è considerato come un insieme unico denominato *gruppo genitore*. Il gruppo genitore viene suddiviso in due sottoinsiemi, individuati a partire dalle modalità (o valori) assunte dalle variabili esplicative: le unità che possiedono un determinato insieme di modalità (o valori), assunte dalle variabili esplicative, costituiscono il primo sottogruppo, le unità complementari il secondo. Ciascuno dei due sottoinsiemi risultanti viene denominato *gruppo figlio*. Il criterio di suddivisione adotta una regola di ottimalità, che cerca di massimizzare l'omogeneità della variabile dipendente entro i gruppi figli, massimizzando al contempo l'eterogeneità tra i gruppi. La regola viene realizzata valutando tutte le suddivisioni possibili rispetto ai valori assunti dalle variabili esplicative e scegliendo la suddivisione che massimizza la regola di ottimalità prescelta.

Nei passi successivi ciascuno dei due gruppi figli del primo passo viene considerato come un gruppo genitore che deve essere ulteriormente suddiviso. Si realizzano quindi ulteriori suddivisioni dei sottoinsiemi ottenuti ai passi precedenti, finché non si verificano le condizioni d'arresto del processo. I criteri di arresto del processo di segmentazione si basano sull'individuazione dell'*albero di taglia minore*, ossia della suddivisione un insieme minimo di gruppi figli, denominati *nodi finali*. Tale insieme deve avere essenzialmente due caratteristiche: *(i)* deve essere facilmente interpretabile e, contestualmente, *(ii)* deve consentire di classificare, nel modo più efficace possibile, le unità statistiche in gruppi che siano strutturalmente diversi rispetto alla variabile dipendente.

L'applicazione di questa tecnica per la valutazione dell'errore sintattico di un archivio amministrativo può essere contestualizzata nel modo seguente:

– il collettivo da sottoporre ad analisi è costituito dai record o dell'intero archivio amministrativo o, alternativamente, del campione opportunamente selezionato;

– la variabile dipendente deve essere una variabile in grado di sintetizzare l'accuratezza sintattica del singolo record. In base a quanto illustrato nei paragrafi precedenti, si può scegliere come variabile dipendente la variabile AS_i definita dalla relazione (3.7). Alternativamente, potrebbe essere utilizzata anche la variabile qualitativa e_i che assume modalità 1,2, o 3 a seconda che il record i-esimo sia corretto o sia soggetto a una violazione debole o forte dell'errore sintattico;

– le variabili esplicative possono essere di due tipi (i) variabili intra-organizzative, che descrivono il processo di acquisizione e trattamento dei dati (ad esempio potrebbe essere utile una variabile che illustra se il singolo indirizzo sia stato acquisito manualmente o tramite una procedura di acquisizione automatica); (ii)

variabili extra-organizzative connesse a peculiari fattori di localizzazione, che possono rendere la georeferenzazione di un indirizzo più facilmente soggetta ad errore sintattico. La scelta di tali variabili viene mostrata nel Paragrafo 3.2.

3.2 Risultati Sperimentali sull'Accuratezza Sintattica

Il presente Paragrafo riporta i risultati di un'analisi sperimentale dell'accuratezza sintattica degli indirizzi sugli archivi, contenenti i dati relativi al 1999, di tre importanti enti pubblici: *(i)* l'INPS, *(ii)* le Camere di Commercio Industria e Artigianato, (*iii*) l'Agenzia delle entrate.

L' archivio INPS (di seguito indicato come *archivio I*) contiene 1.360.133 record, relativi agli indirizzi delle Unità Contributive dell'INPS. Il singolo indirizzo è strutturato nelle seguenti sei componenti: *(i)* sigla della provincia, *(ii)* denominazione del comune, *(iii)* DUG, *(iv)* denominazione dell'area di circolazione, *(v)* numero civico e *(vi)* CAP.

L' archivio anagrafico delle Camere di Commercio Industria e Artigianato (di seguito indicato come *archivio CCIA*) contiene 6.344.449 record relativi agli indirizzi delle unità locali delle imprese. La struttura base del singolo indirizzo prevede le seguenti sei componenti: *(i)* sigla della provincia, *(ii)* denominazione del comune, *(iii)* DUG, *(iv)* denominazione dell'area di circolazione, *(v)* numero civico e *(vi)* CAP. Inoltre, sono presenti due campi aggiuntivi destinati a contenere, rispettivamente, la *località* (o la frazione) ed altre informazioni testuali ritenute utili per localizzare un indirizzo.

L' Archivio anagrafico dell' Agenzia delle Entrate dei soggetti titolari di Partita IVA (di seguito indicato come *archivio AE*) contiene 9.922.701 relativi agli indirizzi di soggetti (fisici e giuridici) titolari di partita IVA. Per ciascun soggetto sono riportati due distinti indirizzi: uno relativo alla sede legale dell'impresa, l'altro relativo al domicilio fiscale. A fini della sperimentazione si è utilizzato solo l'indirizzo della sede legale. La struttura base del singolo indirizzo prevede le seguenti tre componenti: *(i)* codice catastale del comune, *(ii)* punto sul territorio (contenente in un unico campo la DUG, la denominazione dell'area di circolazione e il numero civico), *(iii)* CAP. Si ricorda, che ai fini della valutazione dell'errore sintattico, questa tipologia di formato rende impossibile la misurazione del grado di completezza per le componenti DUG, denominazione dell'area di circolazione e numero civico.

Le principali caratteristiche degli archivi su cui è stata effettuata l'analisi sperimentale dell'accuratezza sintattica sono riassunte in Tabella 3.2.

Denominazione archivio	Numero di record	Variabili toponomastiche disponibili (numero di caratteri riportati in parentesi)	Tipo di unità a cui è riferito l'archivio
Archivio anagrafico dell'INPS (archivio I)	1.360.133	DUG (12) Denominazione area di circolazione (26) Numero civico (5) CAP (5) Sigla provincia (2) Comune (23)	Unità contributive dell'INPS (un'impresa può dare luogo a più unità contributive)
Archivio anagrafico delle Camere di Commercio Industria e Artigianato (archivio CCIA)	6.344.449	Nome comune (30) CAP (5) DUG (3) Denominazione area di circolazione (30) Numero Civico (8) Frazione (25) Altre indicazioni indirizzo (30) Codice Istat comune (3) Sigla provincia (2)	Unità locale (un'impresa può dare luogo a più unità locali)
Archivio anagrafico dell'Agenzia delle Entrate (archivio AE)	9.922.701	SEDE LEGALE Codice catastale comune (4) CAP (5) Punto sul territorio (stringa alfanumerica a formato libero) (35) DOMICILIO FISCALE Codice catastale comune (4) CAP (5) Punto sul territorio (stringa alfanumerica a formato libero) (35)	Soggetti fisici e persone giuridiche titolari di Partita IVA

Tabella 3.2: Schema riassuntivo delle caratteristiche degli archivi su cui è stata misurata l'accuratezza sintattica degli indirizzi

3.2.1 Fasi di Predisposizione delle Basi Dati

Le operazioni preliminari necessarie all'analisi dell'accuratezza sintattica degli archivi sono state: *(i)* la selezione di un campione di record da ciascuno degli archivi; *(ii)* la normalizzazione degli indirizzi campionati; *(iii)* il calcolo degli indicatori di accuratezza. Tali fasi verranno nel seguito brevemente illustrate.

3.2.1.1 Selezione del Campione

Vista la complessità e la mole dei dati dei tre archivi sottoposti a sperimentazione, si è deciso di studiare la loro accuratezza sintattica, analizzando un campione dei record di ciascun archivio. In tal modo, a fronte di un trascurabile errore delle stime, è stato possibile abbattere i costi di elaborazione delle informazioni elementari. Tale campione è stato realizzato adottando il medesimo *disegno di campionamento*[5] per ciascun archivio, questo ha permesso che le analisi finalizzate al confronto tra i diversi archivi non fossero disturbate da differenti tecniche di campionamento.

Il disegno di campionamento adottato per ciascun archivio è analogo a quello illustrato nel Paragrafo 3.1.3.1.1 e può essere sintetizzato nel modo seguente:

- il disegno di campionamento è stratificato con selezione delle unità negli strati senza reimmissione e a probabilità uguali. La variabile di stratificazione è la provincia;
- per ciascun archivio è stato selezionato un campione di 200.000 record. Il numero di record da assegnare a ciascuno strato è stato calcolato mediante la relazione (3.22); il campione estratto in tal modo è autoponderante; questo, nell'ambito di uno specifico archivio, comporta due distinte conseguenze (vedi Paragrafo 3.1.3.1.1): (i) tutti i record hanno approssimativamente la medesima probabilità, pari a n/N, di essere inclusi nel campione (ii) i pesi campionari w_i, con cui ponderare i pesi campionari assumono un valore approssimativamente costante pari a N/n .

Per implementare il disegno sopra descritto, è stato necessario effettuare un'operazione preliminare sui record dell'archivio AE, in quanto la variabile provincia non era presente sui record dell'archivio. Per attuare la procedura di stratificazione provinciale, è stato necessario aggiungere tale variabile a tutti i record dell'archivio, mediante un'operazione di trascodifica del codice catastale del comune.

3.2.1.2 Riconoscimento e Normalizzazione degli Indirizzi Campionati

Come software di riconoscimento normalizzazione è stato utilizzato SISTER [Sister2004] (acronimo SIStema TErritoriale per il Riconoscimento), realizzato negli anni '90 dalla Società SEAT Pagine Gialle sulla base di una convenzione con l'Istat.

Come base dati di indirizzi considerati dal software come indirizzi corretti o di riferimento è stata utilizzata la versione 1999 dello Stradario Nazionale (vedi Paragrafo 3.1.1); si è utilizzata la versione aggiornata al 1999 dello Stradario Nazionale, realizzata dalla società SEAT Pagine Gialle la quale presenta la data di aggiornamento più prossima alla data di riferimento degli archivi amministrativi esaminati nella sperimentazione.

SISTER divide il riconoscimento in due fasi consecutive: *(i)* nella prima, sono esaminate la sigla della provincia e comune che identificano la zona del territorio; *(ii)* nella seconda, qualora la zona sia stata identificata correttamente, SISTER procede al

[5] Con tale locuzione si denota l'insieme delle tecniche adottate e atte a selezionare le unità del campione

riconoscimento e alla normalizzazione del punto sul territorio identificato dalle componenti DUG, denominazione dell'area di circolazione e numero civico.

SISTER può utilizzare cinque diversi livelli di riconoscimento (numerati da 1 a 5), che identificano differenti livelli di rischio di effettuare falsi riconoscimenti (vedi in tal senso il Paragrafo 3.1.3.1.2). Più il livello è elevato, più aumenta la percentuale degli indirizzi riconosciuti in modo erroneo. A livello 1 e 2, il riconoscimento è effettuato per "uguaglianza". Dal livello 3 in poi è utilizzato anche il riconoscimento per "similitudine". Il riconoscimento per uguaglianza richiede che la componente dell'indirizzo in esame (comune o area di circolazione) sia uguale ad una delle denominazioni nota al sistema come denominazione corretta; il riconoscimento per uguaglianza funziona anche con i sinonimi e le forme abbreviate. Il riconoscimento per similitudine richiede che la componente dell'indirizzo da riconoscere possegga un'alta somiglianza con una denominazione, che si presume corretta. Il livello deve essere dichiarato prima dell'inizio del lavoro e si applica sia all'algoritmo di riconoscimento dei comuni, sia all'algoritmo di riconoscimento delle aree di circolazione. Non è possibile utilizzare livelli diversi nei due algoritmi.

Nella sperimentazione è stato utilizzato sempre il livello 3. In base alle sperimentazioni effettuate sui dati si è stabilito, infatti, che tale livello rappresenta un ragionevole compromesso, che permette di *riconoscere e normalizzare* una percentuale sufficiente di indirizzi, senza assumere rischi eccessivi di falsi riconoscimenti.

3.2.1.3 Calcolo degli Indicatori di Accuratezza.

Una volta che il software di riconoscimento e normalizzazione ha restituito i risultati dell'elaborazione, questi sono stati importati in un programma statistico al fine di predisporre le macro e le procedure necessarie per ottenere le misure descritte nel Paragrafo 3.1. Il software statistico utilizzato per la sperimentazione è stato SPSS [SPSS2004]. Esso è sufficientemente diffuso tra gli analisti, tuttavia per lo scopo che ci si prefigge, sono disponibili sul mercato altri software statistici ugualmente versatili.

Utilizzando SISTER si sono dovuti affrontare i problemi (1) e (2), illustrati nel Paragrafo 3.1.3.1.2, connessi al fatto che SISTER: *(i)* prevede un doppio formato: esteso ed abbreviato, per alcune componenti dell'indirizzo normalizzato; *(ii)* per la forma abbreviata, rilascia in un unico campo la DUG e denominazione dell'area di circolazione. Le routine di calcolo adottate per risolvere tali problemi sono quelle riportate nelle figure 3.3 e 3 4. del presente volume.

Relativamente all'archivio AE, si è dovuto affrontare il problema connesso al fatto che il formato dell'indirizzo in tale archivio, non prevede la separazione, in campi distinti delle componenti DUG, denominazione dell'area di circolazione e numero civico. Tale problema non ha permesso il calcolo degli indici di completezza per tale archivio, e per quanto riguarda il calcolo degli indici di accuratezza sintattica è stato risolto utilizzando la metodologia illustrata nel Paragrafo 3.1.3.2 con riferimento al problema (3) di detto Paragrafo.

3.2.2 Principali Risultati sull'Accuratezza Sintattica

Riportiamo ora un'analisi dell'accuratezza sintattica, tale analisi si svilupperà secondo tra livelli di approfondimento. In un primo livello, separatamente per ciascuno degli archivi sottoposti ad indagine, saranno esaminati indici riferiti al complesso dei record dell'archivio (Paragrafo 3.2.2.1). Un secondo livello di analisi esplorerà indici per sottoinsiemi e darà una prima indicazione su alcune possibili associazioni tra grado di accuratezza sintattica e alcune caratteristiche di localizzazione degli indirizzi (Paragrafo 3.2.2.2). Il terzo livello, infine, esplorerà la variabilità osservata dell'accuratezza sintattica, attraverso un modello di segmentazione regressiva (Paragrafo 3.2.2.3).

3.2.2.1 Analisi di Primo Livello sul Complesso dei Record degli Archivi

Sul totale dei record campione, SISTER riconosce una percentuale pari a 87,8%. L'archivio AI è quello che presenta una percentuale minore (86.9%) di indirizzi riconosciuti.

Per quanto riguarda la tipologia di riconoscimento, la quasi totalità dei comuni è stata riconosciuta, da SISTER, per uguaglianza. Per quanto riguarda le componenti che afferiscono all'area di circolazione, quasi la metà è riconosciuta per similitudine.

Tipologia di riconoscimento per componenti dell'indirizzo	Archivi esaminati			Totale archivi
	I	CCIA	AE	
Zona del territorio				
per uguaglianza	85,5	87,7	88,0	87,1
per similitudine	1,3	0,7	0,1	0,7
Aree di circolazione				
per uguaglianza	45,3	49,8	45,3	46,8
per similitudine	41,5	38,6	42,8	41,0
Totale riconosciuti	86,9	88,4	88,1	87,8

Tabella 3.3: Distribuzione percentuale della tipologia di riconoscimento per Archivio

La Tabella 3.4 riporta la distribuzione di frequenza delle motivazioni del non riconoscimento degli indirizzi non identificati da SISTER.

L'errore sintattico più frequente riguarda la denominazione dell'area di circolazione. Tale forma di errore può derivare sia da errori connessi all'acquisizione dei dati mediante procedure di registrazione non controllata, sia da un'insufficiente copertura degli stradari di SISTER. Per questa componente, la percentuale più bassa è quella dell'archivio CCIA (87,5%) e quella più elevata riguarda l'archivio AE.

Per quanto riguarda gli errori sintattici relativi alle località, questi sono stati analizzati caso per caso e si riferiscono ad aggregati sub-comunali (frazioni, borgate, …) non presenti nell'archivio di riferimento di SISTER, oppure a località che non rientrano nel territorio nazionale.

Le aree di circolazione ambigue rappresentano circa il 6,7% del totale degli scarti. Queste riguardano casi dove la coppia DUG e denominazione dell'area di circolazione non contengono informazioni necessarie per identificare un'area di circolazione. Ad esempio, la denominazione "VIA MANZONI" può essere ambigua se, per un dato comune, esiste sia l'area di circolazione VIA MANZONI, sia PIAZZA MANZONI. Infine, con riferimento alle località scartate esse dipendono essenzialmente dai seguenti motivi: *(i)* l'uso di nomi di comuni non più esistenti; *(ii)* denominazioni di comuni non riportati correttamente; *(iii)* denominazioni di frazioni, borgate e altro, non acquisiti nell'archivio di riferimento di SISTER; *(iv)* indirizzi extra-nazionali.

Motivo del non riconoscimento	Archivi esaminati			Totale archivi
	I	CCIA	AE	
Localita' Non Digitata	0,0	0,0	0,0	0,0
Localita' Scartata	0,3	0,1	0,0	0,1
Localita' Ambigua	0,0	0,0	0,0	0,0
Denominazione Area Di Circolazione Scartata	91,1	87,5	91,5	90,1
Civico Via/Cap Non Trovato	0,3	0,3	0,2	0,3
Termini Via Insufficienti	0,2	0,2	0,3	0,2
Denominazione Area Di Circolazione Ambigua	6,9	6,0	7,1	6,7
Denominazione Area Di Circolazione Assente	1,1	6,0	0,9	2,6

Tabella 3.4: Distribuzione di frequenza % per Archivio e motivazione del non riconoscimento dei record non riconosciuti da SISTER

Gli indici *ECCj* (*j*=1,…,5) relativi all'*Errore di Completezza per la Componente j* sono stati ottenuti solo per gli archivi I e CCIA, in quanto per l'archivio AE tali indici non erano calcolabili. Gli indici in oggetto sono riportati nella Tabella 3.5. Come si nota gli errori di completezza sono piuttosto trascurabili, tranne che per la componente numero civico.

Errore di Completezza per Componente	***Archivi esaminati***		***Totale archivi*** (I e CCIA)
	I	***CCIA***	
ECC1 (Errore di completezza per provincia)	0,0001	0,0023	0,0012
ECC2 (Errore di completezza per comune)	0,0000	0,0000	0,0000
ECC3 (Errore di completezza per DUG)	0,0189	0,0496	0,0342
ECC4 (Errore di completezza per area di circolazione)	0,0011	0,0104	0,0058
ECC4 (Errore di completezza per numero civico)	0,1293	0,1492	0,1393

Tabella 3.5: Errore di Completezza per la Componente

La Tabella 3.6 riporta gli indici di *Errore Sintattico per Componente j* (*ESCj*). Dall'analisi della Tabella si desume che l'errore più frequente è relativo alla DUG; un elevato tasso di errore si osserva anche per quanto riguarda la denominazione dell'area di circolazione e il numero civico.

La Tabella 3.7 riporta i valori degli Indici Specifici di Esito per la generica configurazione $_c\phi$ (ISE $_c\phi$).

Si noti che, gli indirizzi corretti, in base ai valori assunti dalla variabile stringa ϕ_i, sono complessivamente maggiori degli indirizzi riconosciuti per uguaglianza da SISTER e riportati nella Tabella 3.2. Quest'apparente discrepanza è imputabile al fatto che gli algoritmi che generano le variabili indicatrici $y_{i,j}$, accertano se ciascuna componente, estesa o abbreviata, costituente l'indirizzo di riferimento sia presente, come sottostringa, nell'indirizzo da valutare, indipendentemente dalla presenza di altre informazioni, corrette o meno, inserite nell'indirizzo stesso. I casi più critici di violazione sintattica riguardano i *pattern* che contengono un 1 nella posizione relativa alla denominazione dell'area di circolazione; ciò può derivare da effettivi errori in fase d'acquisizione dei dati toponomastici, o da un'insufficiente copertura dello Stradario Nazionale utilizzato come riferimento da SISTER. In altri casi dipende da una non corretta o assente acquisizione della DUG.

Errore Sintattico per Componente	Archivi esaminati			Totale archivi
	I	CCIA	AE	
ESC1 (Errore nella Provincia)	0,0010	0,0020	0,0000	0,0010
ESC2 (Errore nel Comune)	0,0290	0,0140	0,0000	0,0140
ESC3 (Errore nella DUG)	0,2100	0,3060	0,2520	0,2560
ESC4 (Errore nella Denominazione area di circolazione)	0,2200	0,1960	0,2130	0,2100
ESC5 (Errore nel Numero civico)	0,2080	0,1800	0,2700	0,2190

Tabella 3.6: Errore Sintattico per Componente

Indicatori Specifici di Esito	Archivi esaminati			Totale archivi
	I	CCIA	AE	
ISE000000	0,6408	0,5862	0,5655	0,5973
ISE000010	0,0514	0,0361	0,1061	0,0645
ISE000100	0,0692	0,0556	0,0622	0,0624
ISE000110	0,0078	0,0046	0,0145	0,0090
ISE001000	0,0551	0,1497	0,0903	0,0984
ISE001010	0,0113	0,0195	0,0254	0,0187
ISE001100	0,0075	0,0156	0,0123	0,0118
ISE001110	0,0021	0,0028	0,0047	0,0032
ISE001111	0,1246	0,1134	0,1189	0,1190
ISE010000	0,0159	0,0083	0,0000	0,0081
ISE010010	0,0024	0,0004	0,0000	0,0009
ISE010100	0,0016	0,0008	0,0000	0,0008
ISE010110	0,0004	0,0001	0,0000	0,0001
ISE011000	0,0015	0,0020	0,0000	0,0012
ISE011010	0,0005	0,0002	0,0000	0,0002
ISE011100	0,0002	0,0002	0,0000	0,0001
ISE011110	0,0001	0,0000	0,0000	0,0000
ISE 011111	0,0064	0,0021	0,0000	0,0028
ISE100000	0,0003	0,0015	0,0000	0,0006
ISE100010	0,0000	0,0000	0,0000	0,0000
ISE100100	0,0000	0,0001	0,0000	0,0000
ISE100110	0,0000	0,0000	0,0000	0,0000
ISE101000	0,0000	0,0003	0,0000	0,0001
ISE101010	0,0000	0,0000	0,0000	0,0000
ISE101100	0,0000	0,0000	0,0000	0,0000
ISE101111	0,0000	0,0003	0,0000	0,0001
ISE110000	0,0000	0,0000	0,0000	0,0000
ISE111010	0,0000	0,0000	0,0000	0,0000
ISE111111	0,0001	0,0001	0,0000	0,0001

Tabella 3.7: Indicatori Specifici di Esito

3.2.2.2 Analisi di Secondo Livello per Sottoinsiemi di Record

La Tabella 3.8 riporta l'indice *AGAS* (Indice Generale di Accuratezza Sintattica) per il dettaglio regionale e complessivo.

Livello territoriale	Archivi esaminati			Totale archivi
	I	CCIA	AE	
Piemonte	0,880	0,843	0,853	0,858
Valle d'Aosta	0,710	0,600	0,656	0,659
Lombardia	0,883	0,885	0,869	0,880
Trentino-Alto Adige	0,674	0,647	0,582	0,638
Veneto	0,788	0,854	0,841	0,827
Friuli-Venezia Giulia	0,843	0,881	0,845	0,856
Liguria	0,866	0,846	0,848	0,853
Emilia Romagna	0,831	0,863	0,840	0,844
Toscana	0,764	0,838	0,812	0,803
Umbria	0,726	0,768	0,729	0,741
Marche	0,801	0,810	0,820	0,810
Lazio	0,883	0,836	0,865	0,861
Abruzzo	0,731	0,740	0,775	0,749
Molise	0,747	0,713	0,763	0,740
Campania	0,814	0,794	0,809	0,805
Puglia	0,818	0,803	0,837	0,820
Basilicata	0,795	0,695	0,763	0,746
Calabria	0,744	0,723	0,746	0,738
Sicilia	0,810	0,804	0,810	0,808
Sardegna	0,796	0,798	0,829	0,808
Totale archivio	0,826	0,827	0,828	0,827

Tabella 3.8: Indice AGAS per Regione e per Archivio

Trasversalmente ai tre archivi, i punteggi non variano in modo rilevante e si assestano intorno a 0,8. Tale invarianza è vera anche a livello regionale, quindi possiamo ragionevolmente affermare che non esistono sostanziali differenze tra i tre archivi studiati.

L'osservazione interregionale del dato medio mostra, invece, differenze più marcate. La regione con il punteggio medio inferiore è il Trentino-Alto Adige, mentre quella con il punteggio medio più elevato è la Lombardia. Per quanto riguarda il Trentino

Alto Adige, il dato non sorprende. Infatti, gli indirizzi che si riferiscono a quest'area geografica presentano un rischio più elevato d'errore sintattico a causa della presenza di una realtà culturale multi-lingue.

Al fine di georeferenziare l'errore sintattico, sono state esplorate le relazioni esistenti tra alcune variabili territoriali, relative al comune cui si riferisce lo specifico record di archivio, e l'appartenenza del record stesso ad un sottoinsieme di indirizzi caratterizzati da una medesima *tipologia d'esito*, distinguendo tra record corretti, record con violazione debole dell'accuratezza sintattica e record con violazione forte dell'accuratezza sintattica (vedi Paragrafo 3.1.1).

Nella Tabella 3.9, per le tre tipologie di esito, sono riportati i valori medi assunti dalle seguenti variabili territoriali:

- densità abitativa per km^2. La densità può essere considerata come una variabile che approssima il grado di conurbazione di un sito;
- numero di zone sub-comunali, riscontrate nell'archivio delle località utilizzato da SISTER;
- distanze in km dei comuni dal capoluogo di provincia.

Dalla Tabella emerge che gli indirizzi con violazione forte ricadono in comuni con densità abitativa tendenzialmente inferiore alla media complessiva. Quest'informazione conforta l'ipotesi che, più l'area amministrativa è scarsamente urbanizzata, maggiore è la possibilità che sia ignorata l'esatta dizione di un indirizzo.Anche per quanto riguarda la relazione tra il numero di località sub-comunali e la persistenza di errori sintattici in un indirizzo sembra che, più elevato è il \grado di frazionamento di un comune, in frazioni e sobborghi, più aumenta la possibilità di riportare dizioni inesatte e non identificabili di un indirizzo.

Tipologia di esito	Archivi esaminati									Totale Archivi		
	I			CCIA			AE					
	Densità	Località Sub Comunali	Distanza Capoluogo	Densità	Località Sub comunali	Distanza Capoluogo	Densità	Località Sub Comunali	Distanza Capoluogo	Densità	Località Sub comunali	Distanza Capoluogo
Corretti	17,0	14,0	18,0	13,6	12,1	21,4	15,3	13,4	20,1	15,4	13,2	19,8
Violazione debole	11,7	11,7	24,3	12,8	11,7	24,6	13,0	12,1	24,3	12,6	11,9	24,4
Violazione grave	8,0	10,8	25,9	7,2	8,6	31,4	9,0	10,6	28,2	8,1	10,0	28,4
Media complessiva	14,6	13,0	20,4	12,7	11,6	23,5	13,8	12,7	22,4	13,7	12,4	22,1

Tabella 3.9: Valori medi per alcune variabili territoriali (densità abitativa per km^2, delle località sub-comunali, della distanza dal capoluogo di provincia) e per tipologia di esito

Rispetto alla distanza dal capoluogo di provincia, i record esenti da errore sono riferiti a comuni che presentano una distanza inferiore alla media complessiva, di converso i record con violazione forte sono localizzati in comuni caratterizzati da una distanza superiore.

La Tabella 3.10 riporta i valori assunti dagli indicatori generali di esito (IGEC, Indicatore Generale di Esito Corretto; IGVD, Indicatore Generale di Violazione debole; IGVF, Indicatore Generale di Violazione Forte) per sottoinsieme di record

relativi a comuni caratterizzati da una medesima zona altimetrica. In tutti e tre archivi il livello di accuratezza sintattica appare crescente al decrescere dell'altimetria.

Archivi	Indicatori Generali di Esito	Montagna interna	Montagna litoranea	Collina interna	Collina litoranea	Pianura	Totale archivio
I	IGEC	0,521	0,634	0,622	0,583	0,691	0,641
	IGVD	0,278	0,220	0,243	0,280	0,197	0,228
	IGVF	0,201	0,146	0,135	0,138	0,113	0,131
CCIA	IGEC	0,485	0,595	0,569	0,529	0,635	0,586
	IGVD	0,311	0,305	0,307	0,334	0,279	0,298
	IGVF	0,204	0,100	0,124	0,137	0,086	0,116
AE	IGEC	0,442	0,428	0,545	0,525	0,616	0,563
	IGVD	0,354	0,459	0,329	0,346	0,290	0,319
	IGVF	0,204	0,113	0,126	0,129	0,094	0,119
Totale archivi	IGEC	0,483	0,550	0,578	0,545	0,648	0,597
	IGVD	0,314	0,330	0,293	0,321	0,254	0,281
	IGVF	0,203	0,120	0,129	0,134	0,098	0,122

Tabella 3.10: Indicatori di esito per zona altimetrica e tipologia di archivio

3.2.2.3 Analisi di Terzo Livello attraverso un Modello di Segmentazione Regressiva

Alla luce delle relazioni emerse in precedenza, è sembrato opportuno esplorare le interazioni tra errore sintattico riscontrato in ciascun record e simultaneamente tutte le caratteristiche territoriali del comune a cui il record appartiene. Infatti, la semplice analisi condotta nel Paragrafo precedente non cattura che aspetti marginali delle relazioni che intercorrono tra tutti i caratteri. Per analizzare l'effetto interattivo delle variabili esplicative territoriali sull'errore sintattico è opportuno adottare strumenti statistici più complessi in grado di apprezzare contestualmente l'insieme delle informazioni presenti nella totalità delle variabili. La tecnica utilizzata si basa su un algoritmo di segmentazione regressiva denominato CART, illustrata nel Paragrafo 3.1.4.

L'applicazione empirica è stata condotta contestualmente sui tre archivi e utilizzando due modelli complementari: il primo impiega come variabile dipendente l'indicatore IEG_i che può assumere tre modalità (indirizzo corretto, indirizzo con violazione debole dell'accuratezza sintattica e indirizzo con violazione forte dell'accuratezza sintattica), il secondo modello utilizza la variabile AS_i, come indicato nel paragrafo 3.1.4.

Per ciascuno dei due modelli sono state considerate i medesimi predittori:

- le variabili territoriali introdotte nel Paragrafo precedente, con l'aggiunta della variabile di stratificazione provincia;
- una variabile che identifica l'archivio amministrativo di riferimento (I, CCIA, AE).

I dendrogrammi riportati in seguito nelle figure 3.6 e 3.7 sintetizzano l'analisi di segmentazione condotta. Per ciascun nodo, le modalità assunte dalle variabili predittrici sono riportate nell'apposita legenda. In entrambi i modelli si sono formati 18 nodi di cui 10 terminali.

Possiamo immediatamente notare che la tipologia di archivio non contribuisce alla suddivisione dei gruppi. Questo dato conferma quanto detto in precedenza, ossia che le modalità procedurali d'aggiornamento e acquisizione degli indirizzi sono qualitativamente omogenee tra le amministrazioni, poiché non sono in grado di produrre sistematiche differenze in termini d'accuratezza conseguita.

L'analisi, dunque, si sposta decisamente sulle caratteristiche territoriali, dove gli indirizzi afferiscono, le quali, per determinati contesti geo-spaziali o culturali, evidenziano alcune persistenze di errori sintattici. La regione e la densità abitativa sono predittori molto potenti e dominano il processo di divisione, escludendo gli altri predittori potenzialmente esplicativi. L'esclusione dipende dal fatto che i predittori sono tra loro correlati. Ad esempio, l'altimetria è correlata con la densità abitativa e porta a far sì che la prima sia esclusa dall'analisi, poiché il predittore associato, la densità, assorbe gran parte della potenzialità esplicativa del primo.

Il nodo 0 rappresenta lo spartiacque, poiché, riporta il valore medio osservato sull'insieme dei tre archivi. Le suddivisioni successive cercano d'identificare gruppi omogenei dalle caratteristiche estreme. I gruppi finali, che presentano un più elevato rischio di presenza di errori sintattici nell'indirizzo, sono rappresentati dai nodi 11, 14 e 17 nel primo modello e 16 e 17 nel secondo modello.

La situazione più critica appare collocarsi nei comuni della provincia di Bolzano, con densità non superiore a 1,8 abitanti per km² (nodo 11 nel primo modello e nodo 16 nel secondo). Tali nodi identificano le aree critiche, ossia sottoinsiemi di record con più elevata probabilità di errore sintattico.

In questo gruppo circa il 74,1% di indirizzi risulta sintatticamente errato è presenta un valore dell'indice AGAS di 0,371, che è molto inferiore al valore medio complessivo.

Il problema della toponomastica della provincia di Bolzano è storico e culturale. La toponomastica ufficiale è tri-lingue: italiana, ladina e tedesca. La questione dei toponimi nasce fondamentalmente durante il fascismo quando il regime, con alcune leggi, abolì i nomi tedeschi, sostituendoli con i nomi italiani. Subito dopo la guerra, alcuni dei nomi originali furono di fatto reintrodotti, ma senza una loro conversione in legge. Esiste una problematica anche per quanto riguarda la toponomastica ladina, ad esempio la località di "Pliscia", la si trova indicata anche come "Plaika", in base ad una derivazione dal tedesco "Plaiken".

I comuni che presentano le situazioni meno problematiche sono identificati nel primo modello dal nodo 5 e nel secondo modello dai nodi 11e 12.

Nel nodo 5 del primo modello si parla di comuni con densità superiore o uguale ad 1,9 abitanti per km², appartenenti alle regioni: Piemonte, Puglia, Lombardia, Lazio, Emilia Romagna, Sardegna, Liguria, Toscana, Veneto, Friuli-Venezia Giulia.

Nel secondo modello ci si riferisce a comuni appartenenti alle regioni: Piemonte, Lombardia, Lazio, Emilia Romagna, Liguria, Friuli-Venezia Giulia, con una densità abitativa, nel nodo 12, non inferiore a 10,1 per km² e nel nodo 11 compresa tra 1,8 e 10,1 per km².

Lo studio effettuato sui tre archivi toponomastici campione consente di generalizzare i risultati ottenuti sui tre archivi studiati, a tutti i comuni italiani che presentano le medesime caratteristiche territoriali. Le tabelle 3.11 e 3.12 sintetizzano queste generalizzazioni, riportando per ciascun nodo terminale, ottenuto con le applicazioni CART, le informazioni territoriali associate. Queste tavole rappresentano un primo immediato strumento di valutazione che può essere facilmente utilizzato da un gestore di dati toponomastici, per identificare nel suo archivio le aree a più elevato rischio d'errore sintattico.

3.3 Valutazione dello Stato Attuale e Prospettive

Un gestore pubblico di dati toponomastici non può prescindere dal valutare il livello di qualità del proprio archivio amministrativo. Tuttavia, in mancanza di software dedicati e di personale professionalmente qualificato, l'ottenimento di questa informazione può essere un'operazione molto complessa da realizzare.

E' cruciale tener presente che, l'implementazione degli indicatori proposti, in questo capitolo, presuppone che a ciascun indirizzo valutato sia stato associato un indirizzo di riferimento ritenuto corretto. Questo processo di associazione è effettuato grazie ad una procedura di riconoscimento e normalizzazione in genere automatizzata. Sul mercato esistono dei software in grado di effettuare questa operazione, tuttavia, occorre tener presente che questi possono produrre dei falsi riconoscimenti. L'indicatore AS_i aiuta ad identificarli. Occorre anche tener presente che non esistendo uno stradario nazionale contenente tutte le aree di circolazione di tutti i comuni italiani, ci si può riferire solo a *database* surrogati che possono non rappresentare completamente la realtà toponomastica ufficiale.

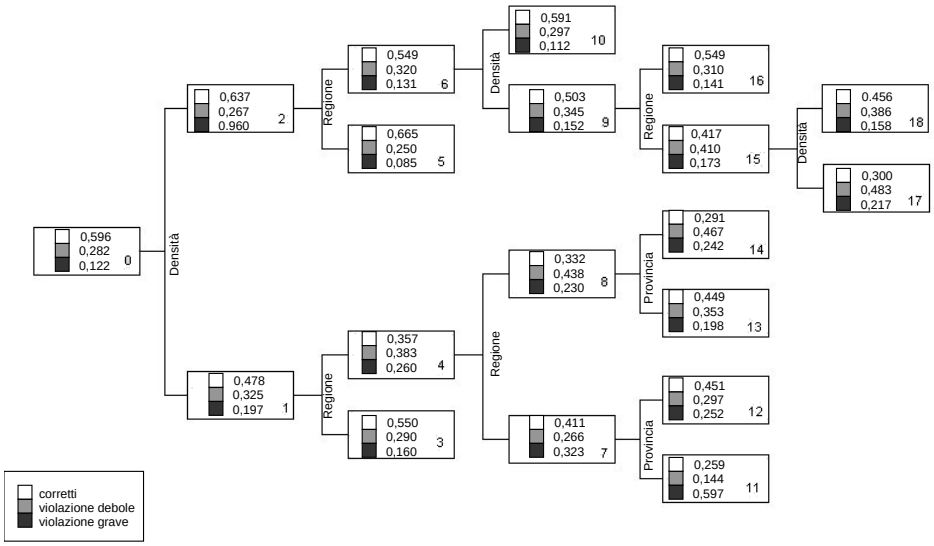
Dal punto di vista dell'analisi, un risultato certamente interessante è che, la peculiarità tipologica dei tre archivi amministrativi non sembra essere un fattore di differenziazione del grado di accuratezza sintattica conseguita, ma che siano piuttosto le caratteristiche geografiche del territorio ad essere i predittori più rilevanti. Ad esempio, i dati toponomastici dei comuni collocati nel Trentino Alto Adige presentano un elevato rischio di errore sintattico, a causa della presenza di una realtà culturale muti-lingua che crea non poca confusione su quale sia la corretta dizione dell'area di circolazione.

Anche nei comuni con un numero elevato di località sub comunali e in quelli poco conurbati il rischio d'errore sintattico degli indirizzi è notevole. In tali casi, spesso, si utilizzano denominazioni d'uso comune, difformi dagli indirizzi ufficiali, oppure incomplete.

In tutte queste circostanze, è utile che il proprietario pubblico dei dati toponomastici si doti di procedure di verifica dell'indirizzo acquisito, al fine di minimizzare il tasso di errore che potrebbe essere introdotto nelle proprie basi di dati.

Dal punto di vista dei possibili sviluppi futuri, l'analisi potrebbe essere affinata valorizzando di più l'elemento geografico di un indirizzo, utilizzando indicatori territoriali ad un livello sub comunale. Le sezioni di censimento costituiscono il livello geografico sub comunale più piccolo con cui sono resi pubblici i dati rilevati dall'Istat e su cui si può attribuire una connotazione spaziale ad un dato toponomastico. Nel censimento del 1991, l'intero territorio nazionale fu suddiviso in 323.395 sezioni. Nel censimento 2001, le sezioni sono quasi raddoppiate e, entro breve tempo, saranno disponibili tutti gli indicatori territoriali utili per poter meglio qualificare il rischio di errore sintattico presente in un archivio di dati toponomastici.

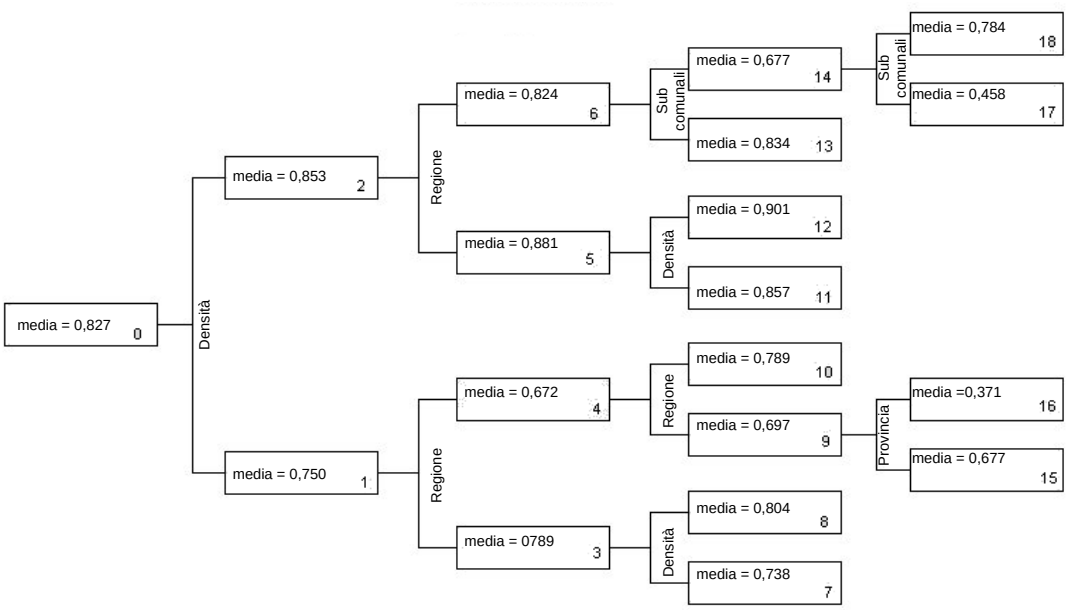
Per georeferenziare l'errore sintattico, di notevole interesse è anche la coppia costituita da comune e CAP, poiché permette di gestire contestualmente una zona del territorio più piccola senza dover scendere nel dettaglio della sezione di censimento. Purtroppo, nel corso degli anni, le aree CAP sono soggette a numerose variazioni e che rendono il loro utilizzo non semplice.



Legenda

Nodo	Percentuale di casi	Variabile discriminante	
		Descrizione	Modalità assunta
1	25,6	Densità	<=1,8830741701266782
2	74,4	Densità	>1,8830741701266782
3	16,0	Regione	Piemonte;Puglia;Lombardia;Lazio;Sicilia;Emilia Romagna;Sardegna; Liguria; Veneto;Friuli-Venezia Giulia
4	9,6	Regione	Toscana;Campania;Abruzzi;Calabria;Marche;Valle d'Aosta; Basilicata;Trentino-Alto Adige;Umbria;Molise
5	56,4	Regione	Piemonte;Puglia;Lombardia;Lazio;Emilia Romagna;Sardegna;Liguria; Toscana;Veneto;Friuli-Venezia Giulia
6	18,0	Regione	Sicilia;Campania;Abruzzi;Calabria;Marche;Valle d'Aosta;Basilicata; Trentino-Alto Adige;Umbria;Molise
7	3,0	Regione	Toscana;Trentino-Alto Adige
8	6,6	Regione	Campania;Abruzzi;Calabria;Marche;Valle d'Aosta;Basilicata;Umbria;Molise
9	8,5	Densità	<=7,8529956773453469
10	9,5	Densità	>7,8529956773453469
11	0,6	Provincia	BZ
12	2,4	Provincia	LI ;LU ;MS ;SI ;PI ;TN ;FI ;PO ;PT ;AR ;GR
13	1,7	Provincia	MT ;AQ ;TR ;AN ;PU ;CB
14	4,9	Provincia	CE ;PE ;RC ;CZ ;MC ;AO ;TE ;SA ;CS ;KR ;CH ;BN ;AV ;AP ;PG ;VV ;PZ ;IS
15	3,0	Regione	Campania;Abruzzi;Calabria;Valle d'Aosta
16	5,5	Regione	Sicilia;Marche;Basilicata;Trentino-Alto Adige;Umbria;Molise
17	0,7	Densità	<=3,0725426567570393
18	2,2	Densità	>3,0725426567570393

Figura 3.6: Dendrogramma dell'analisi di segmentazione binaria mediante CART su 600.551 indirizzi. Variabile dipendente: IEG_i, predittori: regione, provincia, densità abitativa per km^2, numero località sub-comunali, distanza in km dal capoluogo di provincia, altimetria.



Legenda

nodo	Percentuale di casi	Variabile discriminante	
		Descrizione	Modalità assunta
1	25,59	Densità	<=1,8830741701266782
2	74,41	Densità	>1,8830741701266782
3	17,02	Regione	Piemonte;Puglia;Lombardia;Lazio;Sicilia;Emilia Romagna;Sardegna;Liguria;Veneto;Marche;Friuli-Venezia Giulia
4	8,58	Regione	Toscana;Campania;Abruzzi;Calabria;Valle d'Aosta;Basilicata;Trentino-Alto Adige;Umbria;Molise
5	38,38	Regione	Piemonte;Lombardia;Lazio;Emilia Romagna;Liguria;Friuli-Venezia Giulia
6	36,03	Regione	Puglia;Sicilia;Sardegna;Toscana;Campania;Veneto;Abruzzi;Calabria;Marche;Valle d'Aosta;Basilicata;Trentino-Alto Adige;Umbria;Molise
7	3,86	Densità	<=0,58897808593849144
8	13,16	Densità	>0,58897808593849144
9	1,37	Regione	Valle d'Aosta;Trentino-Alto Adige
10	7,20	Regione	Toscana;Campania;Abruzzi;Calabria;Basilicata;Umbria;Molise
11	17,94	Densità	<=10,12932560394
12	20,43	Densità	>10,12932560394
13	33,80	Sub-comunali	<=30
14	2,23	Sub-comunali	>30
15	0,75	Provincia	AO ;TN
16	0,62	Provincia	BZ
17	0,73	Sub-comunali	<=43,5
18	1,49	Sub-comunali	>43,5

Figura 3.7: Dendrogramma dell'analisi di segmentazione binaria mediante CART su 600.551 indirizzi. Variabile dipendente: AS_i , predittori: regione, provincia, densità abitativa per km^2, numero località sub-comunali, distanza in km dal capoluogo di provincia, altimetria.

Nodi	Province	Comuni	Densità	Località sub-comunali	Distanze in km dal capoluogo di provincia	Altimetria	
		Frequenza	Media	Media	Media	carattere modale	Incidenza percentuale
3	AG;AL;AN;AT;BA;BG;BI;BL;BO;BR;BS;CA;CL;CN;CO;CR;CT;EN;FC;FE;FG;FR;GE;GO;IM;LC;LE;LO;LT;ME;MI;MN;MO;NO;NU;OR;PA;PC;PD;PN;PR;PV;RA;RE;RG;RI;RM;RN;RO;SO;SP;SR;SS;SV;TA;TO;TP;TS;TV;UD;VA;VB;VC;VE;VI;VR; VT	3505	0,8	2,9	41,5	1	34,2
5	AL;AR;AT;BA;BG;BI;BL;BO;BR;BS;CA;CN;CO;CR;FC;FE;FG;FI;FR;GE;GO;GR;IM;LC;LE;LI;LO;LT;LU;MI;MN;MO;MS;NO;NU;OR;PC;PD;PI;PN;PO;PR;PT;PV;RA;RE;RI;RM;RN;RO;SI;SO;SP;SS;SV;TA;TO;TS;TV;UD;VA;VB;VC;VE;VI;VR;VT	1969	6,2	3,5	24,1	5	52,4
10	AN;AO;AP;AV;BZ;CB;CE;CH;CS;CT;CZ;EN;MC;ME;NA;PA;PE;PU;RC;RG;SA;SR;TE;TN;VV	199	24,2	3,2	19,2	5	46,7
11	BZ	103	0,6	3,4	51,9	1	100,0
12	AR;FI;GR;LI;LU;MS;PI;PO;PT;SI;TN	398	0,6	4,8	44,2	1	69,1
13	AN;AQ;CB;MT;PU;TR	336	0,6	3,0	49,3	1	51,2
14	AO;AP;AV;BN;CE;CH;CS;CZ;IS;KR;MC;PE;PG;PZ;RC;SA; TE;VV	1153	0,8	2,8	53,5	1	42,0
16	AO;AP;AV;BN;CE;CH;CS;CZ;IS;KR;MC;PE;PG;PZ;RC;SA; TE;VV	208	3,6	3,9	33,8	4	44,7
17	AO;AQ;AV;BN;CE;CH;CS;CZ;NA;PE;RC;SA;TE;VV	110	2,4	2,5	40,2	3	50,0
18	AO;AQ;AV;BN;CE;CH;CS;CZ;KR;NA;PE;RC;SA;TE;VV	125	4,7	3,9	34,5	4	36,8
Totale	Italia	8106	2,8	3,2	38,7	3	31,8

Tabella 3.11: Generalizzazione su tutti i comuni italiani dell'analisi CART ottenuta per la variabile dipendente IEG_i

Nodi	Province	Comuni	Densità	Località sub-comunali	Distanze in km dal capoluogo di provincia	Altimetria	
		Frequenza	Media	Media	Media	carattere modale	Incidenza percentuale
7	AG;AL;AN;AP;AT;BA;BG;BI;BL;BO;BS;CA;CL;CN;CO;CR;CT;EN;FC;FE;FG;FR;GE;GO;IM;LC;LO;LT;MC;ME;MI;MO;NO;NU;OR;PA;PC;PD;PN;PR;PU;PV;RA;RE;RI;RM;RN;RO;SO;SP;SR;SS;SV;TO;TP;UD;VA;VB;VC;VE;VI;VR;VT	1596	0,3	3,1	50,0	1	52,3
8	AG;AL;AN;AP;AT;BA;BG;BI;BL;BO;BR;BS;CA;CL;CN;CO;CR;CT;EN;FC;FE;FG;FR;GE;GO;IM;LC;LE;LO;LT;MC;ME;MI;MN;MO;NO;NU;OR;PA;PC;PD;PN;PR;PU;PV;RA;RE;RG;RI;RM;RN;RO;SO;SP;SR;SS;SV;TA;TO;TP;TS;TV;UD;VA;VB;VC;VE;VI;VR;VT	2097	1,1	2,9	35,4	5	36,7
10	AQ;AR;AV;BN;CB;CE;CH;CS;CZ;FI;GR;IS;KR;LI;LU;MS;MT;PE;PG;PI;PO;PT;PZ;RC;SA;SI;TE;TR;VV	1428	0,7	3,3	53,5	1	43,1
11	AL;AT;BG;BI;BO;BS;CN;CO;CR;FC;FE;FR;GE;GO;IM;LC;LO;LT;MI;MN;MO;NO;PC;PN;PR;PV;RA;RE;RI;RM;RN;SO;SP;SV;TO;TS;UD;VA;VB;VC;VT	1185	4,5	3,2	25,2	5	42,9
12	BG;BI;BO;BS;CO;CR;FC;FR;GE;GO;IM;LC;MI;MO;PN;PV;RM;RN;SO;SP;SV;TO;TS;UD;VA;VB	232	19,6	3,5	16,9	5	61,6
13	AG;AN;AO;AP;AQ;AR;AV;BA;BL;BN;BR;BZ;CA;CB;CE;CH;CL;CS;CT;CZ;EN;FG;FI;GR;IS;KR;LE;LI;LU;MC;ME;MS;MT;NA;NU;OR;PA;PD;PE;PG;PI;PO;PT;PU;PZ;RC;RG;RO;SA;SI;SR;SS;TA;TE;TN;TP;TR;TV;VE;VI;VR;VV	1184	7,4	3,5	28,0	5	43,3
15	AO;TN	271	0,5	2,6	40,6	1	100,0
16	BZ	103	0,6	3,4	51,9	1	100,0
17	LU;TE;VE	3	4,3	37,7	2,0	5	66,7
18	AR;LU;ME;PG;PT;RC	7	5,1	52,1	5,1	1	28,6
Totale	Italia	8106	2,8	3,2	38,7	3	31,8

Tabella 3.12: Generalizzazione su tutti i comuni italiani dell'analisi CART ottenuta per la variabile dipendente AS_i.

4 Miglioramento Interno alle Amministrazioni

Il presente capitolo descrive un insieme di indicazioni che le amministrazioni possono seguire per migliorare la qualità dei dati toponomastici presenti nei propri archivi. Tali indicazioni riguardano due principali aspetti: *(i)* la definizione di un formato standard di acquisizione dei dati toponomastici da parte delle amministrazioni pubbliche; *(ii)* azioni interne alle amministrazioni per il miglioramento della qualità dei dati toponomastici memorizzati.

4.1 I Formati Standard per i Dati Toponomastici

La definizione di formati standard per i dati toponomastici costituisce un elemento fondamentale per controllarne e migliorarne la qualità. La definizione dei formati da adottare consente, infatti, di effettuare controlli di accuratezza, consistenza interna e completezza. Ad esempio, prevedere che una componente dell'indirizzo sia obbligatoria in fase di acquisizione dei dati, consente di limitare i problemi di completezza. Inoltre, qualora siano disponibili dizionari di riferimento per le singole componenti del dato toponomastico, è possibile effettuare controlli di accuratezza.

Nel presente lavoro si distinguono due tipologie di formato per i dati toponomastici: *formato di acquisizione* e *formato di interscambio*.

Il primo formato è relativo alla acquisizione, da parte delle amministrazioni, di dati toponomastici dai soggetti dichiaranti; si distingue in formato di acquisizione *elettronico* e formato di acquisizione *cartaceo*.

Il formato di interscambio è invece relativo alla acquisizione, da parte delle amministrazioni, di dati toponomastici da altre amministrazioni mediante interscambio automatico.

I formati di acquisizione e di interscambio presentano problematiche diverse, essendo differenti i soggetti coinvolti. In particolare, nel formato di acquisizione la rappresentazione delle componenti dell'indirizzo dovrebbe preferibilmente essere espressa in linguaggio naturale, in virtù dell'interazione con gli utenti. Viceversa, nel formato di interscambio, è auspicabile l'utilizzo di codici numerici per l'identificazione dei campi scambiati, al fine di minimizzare la quantità di errori connessi alla trasmissione ed al trattamento di stringhe testuali. Inoltre, l'utilizzo di codici oltre a favorire l'univocità dell'informazione scambiata, favorisce una compattezza nell'interscambio che aumenta l'efficienza della trasmissione. Si evidenzia che l'utilizzo di codici nell'interscambio non pregiudica la comprensione dell'informazione da parte di utenti umani, qualora si prevedano opportuni dizionari di conversione. Alcuni di questi dizionari sono stati opportunamente progettati nell'ambito del presente lavoro, e vengono descritti nel seguito del presente capitolo ed allegati in Appendice.

Per l'acquisizione elettronica e per l'interscambio si è assunto di utilizzare file XML [XML2004]. L'utilizzo di XML è motivato dall'esigenza di aderire ad un formato ampiamente adottato e non vincolante, con riferimento a specifiche tecnologie di supporto.

In Figura 4.1, sono rappresentati entrambi i formati e i soggetti coinvolti nell'acquisizione ed interscambio dei dati toponomastici.

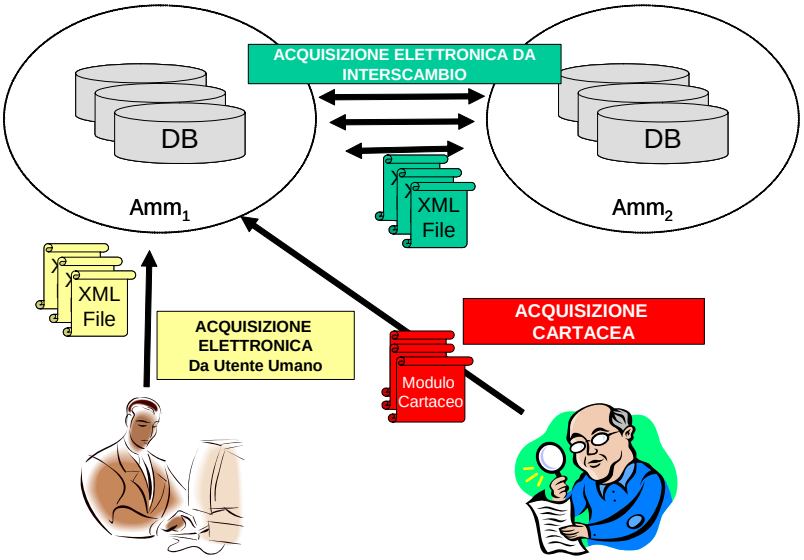


Figura 4.1: I formati di acquisizione ed interscambio dei dati toponomastici.

Un aspetto estremamente importante è relativo al formato di memorizzazione dei dati toponomastici nelle basi di dati delle amministrazioni. Si è effettuata la scelta progettuale di non imporre alle amministrazioni un formato di memorizzazione di tipo specifico per i dati toponomastici. Si è deciso, invece, di regolamentare l'interfacciamento di ciascuna amministrazione verso gli utenti e verso le altre amministrazioni, mediante la definizione dei formati standard di acquisizione e di interscambio. L'adozione di questi formati comporta che un'amministrazione intraprenda, con riferimento alla memorizzazione dei dati toponomastici, una delle seguenti attività:

– adeguamento delle proprie basi di dati ai formati di acquisizione ed interscambio;
– adozione di strumenti di conversione dai formati proposti al formato di memorizzazione adottato da ciascuna amministrazione.

Entrambe le scelte comportano investimenti economici da parte delle singole amministrazioni.

La prima scelta risulta essere la più costosa in quanto prevede la riprogettazione logica e fisica delle basi di dati che memorizzano l'informazione toponomastica ed operazioni di normalizzazione per popolare la nuova base di dati. Ad esempio, laddove in alcune basi di dati l'indirizzo è memorizzato in un'unica stringa, si rende necessaria un'operazione di normalizzazione rispetto alle componenti costitutive. Inoltre, le procedure applicative di gestione dei dati devono ulteriormente essere modificate, al fine di interfacciarsi correttamente con le nuove basi di dati. Tuttavia, tale scelta è sicuramente preferibile sul lungo termine, in quanto l'alimentazione delle basi di dati toponomastiche può avvenire in maniera diretta, a partire dai processi di acquisizione ed interscambio. Inoltre, le amministrazioni possono più facilmente inserirsi nel circuito di cooperazione applicativa, mediante l'adozione di standard per

l'interoperabilità semantica e tecnologica a tutti i livelli di gestione dei dati applicativi.

La seconda scelta può essere adottata sul breve termine, di modo da consentire comunque alle amministrazioni di interoperare mediante l'utilizzo di formati standard nell'acquisizione e nell'interscambio. Da un punto di vista economico, l'investimento è più limitato e dunque maggiormente gestibile anche da amministrazioni di dimensioni più limitate.

Nel seguito del presente capitolo ci si focalizzerà sulla descrizione del formato di acquisizione. Mentre, nel Capitolo 5, sarà descritto il formato di interscambio.

4.2 Il Formato di Acquisizione

La definizione del formato di acquisizione ha comportato la preliminare individuazione di un insieme di regole guida che hanno portato alla proposta del formato stesso. Sulla base di tali regole sono state individuate le componenti del dato toponomastico e la loro rappresentazione nell'ambito del formato di acquisizione. Inoltre, per l'acquisizione elettronica sono state progettate delle soluzioni per effettuare controlli di accuratezza sintattica e di completezza sui dati dichiarati dai soggetti prima che tali dati vengano memorizzati negli archivi interni delle amministrazioni. In particolare, tali controlli sono realizzati mediante l'utilizzo congiunto di dizionari di riferimento e di uno schema XML che implementa il formato proposto. Il linguaggio utilizzato per la definizione dello schema è XML Schema [XMLSchema1][XMLSchema2] che risulta, allo stato attuale, il linguaggio che si sta affermando come standard riconosciuto dal World Wide Web Consortium (W3C).

Nel Paragrafo 4.2.1 sono descritti i criteri di individuazione delle componenti di un indirizzo; nel Paragrafo 4.2.2 sono illustrati i formati proposti per le componenti; infine, nel Paragrafo 4.2.3 le soluzioni progettate per i controlli di qualità sono presentate nel dettaglio.

4.2.1 Le Regole per l'Individuazione delle Componenti

La definizione di un formato standard per l'indirizzo prevede che venga individuato un insieme di componenti, che si possa affermare come riferimento riconosciuto dai principali soggetti che memorizzano informazione toponomastica.

A tal fine, l'insieme selezionato delle componenti deve godere delle seguenti proprietà:

1. avere un fondamento normativo,
2. rispettare la struttura gerarchica del dato toponomastico,
3. evitare ridondanze e consentire di caratterizzare l'indirizzo mediante un insieme minimo di componenti.

Alla luce dell'analisi effettuata sulle componenti degli indirizzi in alcuni archivi amministrativi, sintetizzata nella Tabella 2.1 del Capitolo 2, gli elementi che possono essere candidati alla costituzione dell'indirizzo stesso, sono: la provincia, il comune, il

codice di avviamento postale, la località/frazione, la DUG, la denominazione dell' area di circolazione e il numero civico.

Ciascuna delle tre proprietà elencate sarà considerata in seguito con riferimento alle componenti del dato toponomastico.

4.2.1.1 Fondamento Normativo

Il presente volume riporta un'appendice sulla normativa in materia di localizzazione territoriale delle persone fisiche e delle unità economiche, in cui è riportata un'approfondita descrizione della situazione legislativa vigente. Si ritiene tuttavia opportuno sintetizzare alcuni principali elementi normativi che hanno guidato le scelte effettuate sul formato. In particolare, i principali riferimenti normativi sono la Costituzione Italiana, le leggi regionali, il Regolamento Anagrafico dei comuni [ISTAT1992] e il relativo regolamento di attuazione approvato con D.P.R. n. 136/1958 e sostituito dal DPR n. 223/1989.

Tali riferimenti aiutano a identificare le specifiche componenti del dato toponomastico. Per alcune componenti la normativa vigente è chiara e definita; comune e provincia sono infatti definite nella Costituzione Italiana. L'elenco ufficiale dei comuni e delle province italiane è altresì disponibile sul sito http://www.istat.it/.

Il Regolamento Anagrafico definisce la struttura delle aree di circolazione. In particolare, è demandato all'Istat il compito di definire le norme tecniche che regolamentano l'esecuzione degli adempimenti dei Comuni in materia topografica ed ecografica, ivi compresa l'assegnazione delle nomenclature alle aree di circolazione e delle numerazioni degli accessi da queste alle unità ecografiche. Secondo il D.P.R. n. 223/1989, a livello locale, è compito del Comune che ne è competente per territorio assegnare ad ogni area di circolazione una propria distinta denominazione. Nell'ambito delle istruzioni, per la formazione delle basi territoriali e per l'ordinamento ecografico, viene fatta una distinzione tra aree di circolazione all'interno dei centri abitati (cfr. Glossario), dotati di regolare rete stradale o meno, oppure esterne ai centri stessi.

Nel caso di aree di circolazione interne ai centri abitati, ogni spazio del suolo pubblico o aperto al pubblico di qualsiasi forma e misura, destinato alla viabilità, costituisce una separata area di circolazione che, in quanto tale, deve essere distinta con una propria denominazione. Costituisce una distinta area di circolazione ogni via, strada, corso, viale, vicolo, calle, salita, piazza, piazzale, largo, campiello e simili, comprese le strade private purché aperte al pubblico.

Nel caso di aree di circolazione al di fuori dei centri abitati, tutti gli spazi destinati alla viabilità sono costituiti dal complesso delle strade, stradoni, carrarecce, mulattiere, sentieri e simili, che collegano i centri abitati con i nuclei e le case sparse che gravitano su di essi. In questo caso, la denominazione di ogni area di circolazione deve essere la stessa della rispettiva località.

Il Regolamento Anagrafico vieta esplicitamente che una stessa denominazione sia data a più aree di circolazione della stessa specie, mentre è ammessa l'omonimia, se si tratta di aree di circolazione di specie diversa. La *specie* di una area di circolazione corrisponde esattamente al concetto di DUG.

Sempre nel regolamento anagrafico viene introdotta la *numerazione civica*. Questa è costituita dai numeri che contrassegnano gli accessi esterni, cioè quelli che dall'area di circolazione immettono, direttamente o indirettamente, alle unità ecografiche. Ogni area di circolazione deve avere una propria numerazione civica, che può essere ordinata o secondo la successione naturale dei numeri o secondo il sistema metrico.

Dall' analisi della normativa emerge quindi che comune e provincia sono definiti ufficialmente e pubblicati. Inoltre, con riferimento all'area di circolazione, la normativa definisce tre componenti: la DUG, la denominazione, e il numero civico.

Le altre componenti del dato toponomastico, quali il CAP e la località, non sono soggette a normative specifiche. Il CAP è tuttavia ufficialmente definito e gestito da Poste Italiane.

4.2.1.2 Struttura Gerarchica del Dato Toponomastico

Alcune componenti del dato toponomastico possono essere organizzate sulla base di una gerarchia a livelli, tale per cui l'elemento a livello i+1 *contiene* l'elemento a livello i. Tali componenti includono provincia, comune, area di circolazione (denominazione e DUG) e numero civico; la gerarchia è mostrata in Figura 4.2.

Sia il codice di avviamento postale che la località/frazione non si inseriscono ad un unico livello di tale struttura gerarchica. Ad esempio, una località/frazione può essere subcomunale (e.g. località Istat) o anche sovra-comunale (e.g. località postale in alcuni casi).

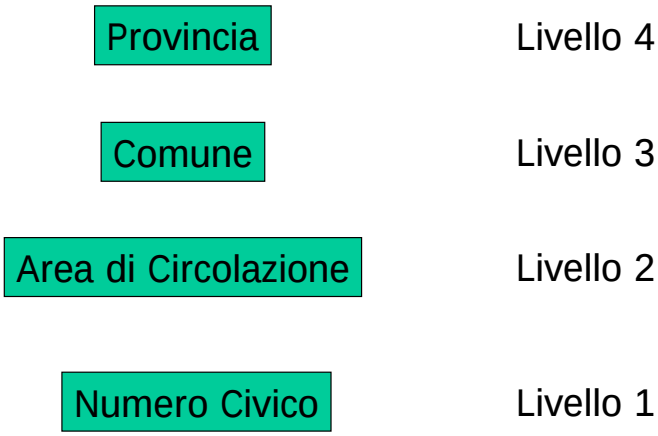


Figura 4.2: Struttura gerarchica del dato toponomastico

4.2.1.3 Ridondanza e Minimalità delle Componenti

Tra le componenti del dato toponomastico risulta sempre ridondante il CAP. Infatti, il CAP contiene informazioni sulla coppia comune e provincia e nei grandi comuni anche sull'area di circolazione.

Sulla base della normativa esistente, la località non è necessaria per identificare univocamente l'indirizzo. Infatti, come riportato sopra, il Regolamento Anagrafico vieta l'omonimia di aree di circolazione della medesima specie e quindi, qualora sia nota l'area di circolazione, non risulta necessario specificare anche la località. Tuttavia, sulla base delle analisi empiriche effettuate, si sono riscontrate alcune rare eccezioni alle norme, previste dal Regolamento Anagrafico, con casi di omonimia delle aree di circolazione distinguibili solo in base alla località. Tali eccezioni si verificano nel caso di fusione di più comuni, a seguito delle quali non avviene una ridenominazione delle aree di circolazione.

4.2.2 L' Insieme delle Componenti

Dall'analisi empirica effettuata sugli archivi della Pubblica Amministrazione e dalle considerazioni sopra descritte risulta che tutte le componenti del dato toponomastico esaminate rivestono un ruolo importante.

Tuttavia, ai fini della definizione di un formato standard, sia di acquisizione che di interscambio, si ritiene opportuno suddividere dette componenti in *obbligatorie* ed *opzionali*.

L'obbligatorietà è legata all' individuazione univoca sul territorio di una unità ecografica e/o di un soggetto fisico o giuridico a questa associata; ad esempio, l'assenza della componente provincia può inficiare l'identificazione univoca dell'indirizzo, qualora due comuni con la stessa denominazione appartengano a province diverse. Le componenti obbligatorie devono dunque essere sempre presenti negli indirizzi, mentre le componenti opzionali possono non essere presenti.

Le *componenti* che si propongono come *obbligatorie* sono: provincia, comune, DUG, denominazione dell' area di circolazione e numero civico. Infatti, dalla normativa e dalla struttura gerarchica del dato toponomastico, discende la possibilità di identificare univocamente l'indirizzo mediante tutte e sole tali componenti.

Le *componenti opzionali* risultano: località/frazione e CAP. Infatti, il CAP è ridondante e non esplicitamente normato, e la località/frazione risulta non essenziale ai fini dell'identificazione univoca dell'indirizzo, almeno secondo la normativa e nella maggioranza dei casi concreti. Tuttavia, sia il CAP che la località/frazione sono state incluse nell'insieme delle componenti costitutive dell'indirizzo, in quanto possono risultare utili in una pluralità di situazioni. Ad esempio, la località può risultare utile per risolvere le eccezioni di omonimia delle aree di circolazione a livello comunale; il CAP è utile per il recapito postale.

4.2.3 Il Formato delle Componenti

Una volta definite le componenti del dato toponomastico, ove in particolare sono stati distinti gli elementi in obbligatori e opzionali, è necessario stabilire il formato di rappresentazione di tali componenti.

Nel presente paragrafo saranno trattate singolarmente le varie componenti, considerando dapprima quelle obbligatorie e successivamente quelle opzionali.

Si ipotizza l'acquisizione del dato in un archivio nazionale, in cui il livello territoriale di riferimento è lo Stato Italiano. Per tale motivo viene omesso l'eventuale campo stato/nazione, che dovrebbe invece essere presente in un eventuale contesto internazionale.

Per ciascuna componente sarà considerata la sua rappresentazione sia nell'ambito del formato di acquisizione cartaceo, che nell'ambito del formato di acquisizione elettronico.

Le Componenti Del Dato Toponomastico	
Componenti Obbligatorie	Componenti Opzionali
Provincia	
Comune	CAP
DUG	Località/Frazione
Denominazione Area di Circolazione	
Numero Civico	

Tabella 4.1- Componenti dell'indirizzo e relative obbligatorietà

4.2.3.1 Componenti Obbligatorie

4.2.3.1.1 Provincia

La rappresentazione che si suggerisce di utilizzare per la componente *provincia* è la *sigla automobilistica della stessa*.

Nell' acquisizione delle informazioni ad opera di un soggetto esterno che ne dichiara l'ingresso in un sistema o in un archivio, non è necessario né previsto che invece della sigla venga utilizzato il nome della provincia per esteso o un codice numerico, come ad esempio il codice Istat di tre cifre numeriche. Infatti, la sigla della provincia è stata preferita al nome per esteso della provincia, in quanto induce minori errori di immissione; inoltre, risulta un campo tipicamente conosciuto dal soggetto dichiarante. La sigla automobilistica è altresì la modalità di memorizzazione della componente provincia nella maggioranza delle basi di dati pubbliche, per cui non risulta necessario prevedere alcun livello di conversione intermedia tra l'acquisizione e la memorizzazione. Infine, la sigla automobilistica si è preferita ad un codice in virtù della maggiore interpretabilità da parte dei soggetti dichiaranti.

Formato di Acquisizione Cartaceo

L'acquisizione cartacea della componente *provincia* prevede la presenza di un campo di due caratteri.

Formato di Acquisizione Elettronico

L'acquisizione elettronica della componente *provincia* prevede che il campo provincia sia di un tipo enumerato definito dall' insieme delle sigle delle province ammissibili. Questa soluzione consente di controllare i valori immessi, riducendo errori di accuratezza dovuti, ad esempio, a digitazioni errate.

4.2.3.1.2 Comune

La rappresentazione che si suggerisce di utilizzare per la componente *comune* è il *nome del comune riportato per esteso*.

Una scelta alternativa poteva prevedere l'utilizzo dei Codici Comune Istat o Codici Comune del Ministero dell'Economia e delle Finanze; tale scelta è stata scartata per i motivi, sopra esposti, di scarsa interpretabilità dei codici.

Formato di Acquisizione Cartaceo

L'acquisizione cartacea della componente *comune* prevede la presenza di un campo di almeno quaranta caratteri. La rappresentazione tramite quaranta caratteri è sufficientemente lunga da consentire che il nome del comune, anche se composto di più parole, possa essere sempre contenuto in un unico campo.

E' opportuno che il nome del comune sia riportato il più possibile per esteso, onde evitare abbreviazioni che possano contribuire ad ingenerare confusione ai fini della corretta identificazione.

Per il principio della completezza, è importante anche che siano riportate tutte le parole di un nome composto.

Ad esempio, non sono ammesse le dizioni "S. Benedetto" o "S. Benedetto del Tron." invece di "San Benedetto del Tronto". Tuttavia, anche le notazioni abbreviate potrebbero essere gestite, in una successiva memorizzazione su supporto elettronico, in virtù del fatto che, in alcuni archivi, sono registrate le più frequenti abbreviazioni e permutazioni delle dizioni ufficiali.

Formato di Acquisizione Elettronico

L'acquisizione elettronica della componente *comune* prevede che il campo comune sia di un tipo enumerato definito dall' insieme delle denominazioni ufficiali dei comuni. Le nomenclature ufficiali dei comuni sono rilevate e archiviate dall'Istat, che già le rende pubbliche sul proprio sito. Sarà questo archivio di denominazioni ufficiali che sarà preso come archivio di riferimento (vedi Dizionario dei Comuni in Appendice).

4.2.3.1.3 Area di Circolazione

La componente relativa all'area di circolazione è composta da due elementi, la DUG (ad esempio, Via/Piazza) e la denominazione vera e propria. Si prevede che questi elementi siano acquisiti in due campi differenti.

Ad esempio, l'area di circolazione "Via Leonardo Da Vinci", deve essere distinta nelle due componenti elementari:

- DUG: Via
- Denominazione: Leonardo Da Vinci

DUG

Formato di Acquisizione Cartaceo

L'acquisizione cartacea della componente *DUG* prevede la presenza di un campo di venticinque caratteri.

Come criterio generale, si devono includere nella DUG anche gli eventuali complementi alla DUG stessa, come le preposizioni o qualsiasi altra particella che non sia strettamente legata alla denominazione (ad esempio, Via dei o Via della). E' possibile che la distinzione delle componenti DUG e denominazione di un'area di circolazione non sia sempre così intuitivo, come nell'esempio riportato sopra. Tale difficoltà è risolta nel formato di acquisizione elettronico, che si avvale di un specifico vocabolario delle DUG. La eventuale consultazione di tale vocabolario anche in fase di acquisizione cartacea, potrebbe essere di ausilio alla corretta identificazione delle DUG. Come ulteriore regola per l'identificazione corretta delle DUG, si suggerisce di individuare in primo luogo la denominazione dell'area di circolazione e successivamente, identificare la componente DUG con la parte restante dell'area di circolazione.

Formato di Acquisizione Elettronico

L'acquisizione elettronica della componente DUG prevede che il campo ad essa relativo sia di un tipo enumerato, definito dall' insieme delle DUG ammissibili, presenti nel dizionario di riferimento delle DUG. Tale dizionario, riportato in Appendice e descritto nel successivo paragrafo 4.3.1.1, è il risultato dell'integrazione dei dizionari delle DUG gestiti da ISTAT e Poste Italiane.

Denominazione

Formato di Acquisizione Cartaceo

L'acquisizione cartacea della componente *denominazione* prevede la presenza di un campo di quaranta caratteri, dunque una stringa sufficientemente lunga come per tutti i campi che contengono una denominazione.

Formato di Acquisizione Elettronico

L'acquisizione elettronica della componente *denominazione* prevede che il campo denominazione sia di tipo *stringa*.

E' realistico prevedere che in futuro venga realizzato un dizionario delle strade a livello nazionale, che possa definire l'elenco delle vie ufficiali esistenti su tutto il territorio nazionale, e un dizionario delle strade per ciascun comune, ovvero l'elenco delle strade conosciute per singolo comune. In tale caso anche il formato del campo denominazione potrebbe essere di tipo enumerato sui valori di tale dizionario.

4.2.3.1.4 Numero Civico

Il numero civico è una componente articolata del dato toponomastico. Essa è rappresentata in maniera differente, nei vari archivi analizzati e spesso non è sufficientemente specificata. Si riporta, in Tabella 4.2, un'analisi delle diverse rappresentazioni del numero civico negli archivi analizzati,

Archivio	Componenti	Note
Istat	– Parte Numerica – Parte Alfanumerica/Esponente – Tipologia Parte Numerica – Rosso/Nero	
Poste Italiane	– TipoCivico – CivicoNumerico – CivicoLetterale – Esponente – NeroRosso – Km	– Il TipoCivico può essere P=Principale, D=Derivato e K=chilometrico. – CivicoLetterale rappresenta il civico quando è letterale., – Esponente in caso di derivato. – NeroRosso se rosso vale R. – Km rappresenta il civico quando è chilometrico.
Seat	– Parte Numerica – Parte Alfanumerica/Esponente – Tipologia Parte Numerica – Rosso/Nero	
Camere di Commercio	Unico campo senza distinzione in componenti	
Agenzia delle Entrate	Il numero civico non costituisce un campo separato, ma è incluso in una stringa "indirizzo"	
Inps	Unico campo senza distinzione in componenti	

Tabella 4.2: Le componenti del numero civico negli archivi analizzati.

al fine di motivare la scelta effettuata per la rappresentazione di tale componente nei formati standard.

Come evidenziato dalla Tabella 4.2, la rappresentazione del numero civico è eterogenea nei vari archivi; infatti, dalla rappresentazione come campo unico di INPS e Camere di Commercio, si passa alla rappresentazione estremamente dettagliata di Poste Italiane.

Per rappresentare il numero civico nei formati standard, si è scelto di considerare le seguenti componenti:

- Parte Numerica
- Tipologia Parte Numerica
- AlfaNumerico
- Rosso Nero
- Civico Assente

La scelta effettuata prevede l'adozione dell'insieme più dettagliato di componenti del numero civico, ovvero quello adottato da Poste Italiane, con alcune semplificazioni. In particolare, le componenti KM e Civico Numerico di Poste Italiane (vedi Tabella 4.2) sono state unificate nella componente ParteNumerica. Inoltre, la componente di Poste Italiane, relativa al Civico Letterale, non è stata inclusa in quanto la presenza di unità ecografiche con un civico letterale è estremamente limitata e non è prevista dalla normativa.

Il numero civico nel suo complesso è obbligatorio. L'obbligatorietà delle singole componenti è rappresentata in Tabella 4.3.

	Parte numerica	Tipologia parte numerica	Alfanumerico	RossoNero	CivicoAssente
Unità ecografica senza numero civico	Nessun valore deve essere specificato	Nessun valore deve essere specificato	Nessun valore deve essere specificato	Nessun valore deve essere specificato	Obbligatorio
Unità ecografica con numero civico	Obbligatoria	Obbligatoria	Opzionale	Opzionale	Nessun valore deve essere specificato

Tabella 4.3: Obbligatorietà delle componenti del numero civico.

Parte Numerica

La componente *parte numerica* è costituita o da un progressivo numerico, oppure da una distanza. Nei formati standard si sceglie di esprimere la distanza in metri anziché, come accade in alcuni archivi, rappresentarla in chilometri. Tale scelta è fondata sulla necessità di semplificarene l'acquisizione. Infatti, esprimere la distanza in chilometri può richiedere di introdurre cifre decimali e dunque di specificare una virgola; questa complicazione, nella specifica della distanza espressa in chilometri, può ingenerare errori.

Formato di Acquisizione Cartaceo

L'acquisizione cartacea della componente *parte numerica* prevede la presenza di un campo di sei caratteri.

Formato di Acquisizione Elettronico

L'acquisizione elettronica della componente parte numerica prevede che il campo parte numerica sia di tipo *intero*.

Tipologia Parte Numerica

Al fine di distinguere quando la componente numerica rappresenta un progressivo numerico, oppure una distanza, si è scelto di introdurre un' ulteriore componente

che ne specifica la natura, cioè la componente *tipologia parte numerica*. La tipologia parte numerica assume il valore *m* (metri), quando la parte numerica esprime una distanza; assume, invece, il valore *n* (numero) quando la parte numerica è un progressivo numerico. Tale campo è obbligatorio se è presente un valore per il campo parte numerica (vedi Tabella 4.3).

Formato di Acquisizione Cartaceo

L'acquisizione cartacea della componente *tipologia parte numerica* prevede la presenza di un campo di un unico carattere.

Formato di Acquisizione Elettronico

L'acquisizione elettronica della componente *tipologia parte numerica* prevede che il campo ad essa relativo sia di tipo enumerato, con due valori ammissibili corrispondenti ad *m* (metri), quando la parte numerica esprime una distanza ed *n* (numero) altrimenti.

Alfanumerico

La parte alfanumerica è convenzionalmente definita "esponente". Un esempio di numero civico con parte alfanumerica è *45 B/4* , dove *45* rappresenta la parte numerica e *B/4* rappresenta la parte alfanumerica o esponente. Se esistono dei caratteri separatori (e.g., /, -) rispetto alla parte numerica, essi sono da includere nella parte alfanumerica. Ad esempio, il numero civico *45/B* si suddivide nella parte numerica *45* e nella parte alfanumerica */B*. Il campo è opzionale, anche se è presente la parte numerica (vedi Tabella 4.3).

Formato di Acquisizione Cartaceo

L'acquisizione cartacea della componente alfanumerica prevede la presenza di un campo di 20 caratteri.

Formato di Acquisizione Elettronico

L'acquisizione elettronica della componente alfanumerica prevede che il campo alfanumerico sia di tipo *stringa*.

RossoNero

Il campo RossoNero è utilizzato per specificare ulteriormente la numerazione civica in alcuni comuni italiani, in cui la distinzione tra rosso e nero consente di distinguere tra unità ecografiche con finalità commerciali e non. Il campo RossoNero può assumere i valori *r* (rosso) e *n* (nero). Il campo è opzionale (vedi Tabella 4.3).

Formato di Acquisizione Cartaceo

L'acquisizione cartacea della componente *RossoNero* prevede la presenza di un campo di un unico carattere.

Formato di Acquisizione Elettronico

L'acquisizione elettronica della componente *RossoNero* prevede che il campo corrispondente sia di tipo enumerato con due valori ammissibili corrispondenti ad *r* (rosso) e *n* (nero).

CivicoAssente

Il numero civico è un campo obbligatorio per l'esatta individuazione della localizzazione di un soggetto. Tuttavia, un'area di circolazione può non essere dotata di una propria numerazione civica; in tali casi la componente *CivicoAssente* assume il valore convenzionale *snc* (senza numero civico). Il campo è obbligatorio qualora non sia specificato un valore per la componente parte numerica (vedi Tabella 4.3). L'introduzione di tale regola di obbligatorietà consente di distinguere errori di incompletezza, nella specifica del numero civico (il numero civico esiste ma non è specificato dal dichiarante), dai casi in cui il numero civico è effettivamente assente.

Formato di Acquisizione Cartaceo

L'acquisizione cartacea della componente *CivicoAssente* prevede la presenza di un campo di tre caratteri.

Formato di Acquisizione Elettronico

L'acquisizione elettronica della componente CivicoAssente prevede che il campo relativi, se specificato, assuma il valore di *default* "snc", altrimenti tale valore non deve essere specificato.

4.2.3.2 Componenti Non Obbligatorie

4.2.3.2.1 Codice di Avviamento Postale

Il CAP è gestito da Poste Italiane ed è attualmente costituito da 5 cifre numeriche. Tale composizione del CAP potrebbe essere rivista alla conclusione del progetto di Poste Italiane, in cui si prevede la nuova numerazione dei CAP come una rappresentazione di 9 cifre numeriche.

Formato di Acquisizione Cartaceo

L'acquisizione cartacea della componente *CAP* prevede la presenza di un campo di cinque caratteri.

Formato di Acquisizione Elettronico

L'acquisizione elettronica della componente CAP prevede che il campo CAP sia di un tipo enumerato, i cui valori corrispondono all'elenco attuale dei CAP italiani (vedi Appendice).

4.2.3.2.2 Località/frazione

Formato di Acquisizione Cartaceo

L'acquisizione cartacea della componente *località* prevede la presenza di un campo di 40 caratteri.

Formato di Acquisizione Elettronico

L'acquisizione elettronica della componente *località* prevede che il campo località sia di un tipo enumerato, i cui valori corrispondono ad un dizionario di riferimento per le località italiane (si veda Paragrafo 4.2.3.1).

4.2.4 Il Formato di Acquisizione Elettronico

Il formato di acquisizione elettronico è definito dall'XML Schema rappresentato in Figura 4.3.

```
<?xml version="1.0" encoding="UTF-8"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema">
  <xs:annotation>
    <xs:documentation>Questo file contiene la definizione del formato di acqusizione elettronico di ciascuna componente del dato toponomastico</xs:documentation>
  </xs:annotation>
  <xs:include schemaLocation="http://www.istat.it/.../ListaProvince.xsd"/>
  <xs:include schemaLocation=" http://www.istat.it/.../ListaComuni.xsd"/>
  <xs:include schemaLocation=" http://www.istat.it/.../ListaLocalita.xsd"/>
  <xs:include schemaLocation=" http://www.istat.it/.../ListaCAP.xsd"/>
  <xs:include schemaLocation=" http://www.istat.it/.../DenominazioneType.xsd"/>
  <xs:include schemaLocation=" http://www.istat.it/.../ListaDUG.xsd"/>
  <xs:include schemaLocation=" http://www.istat.it/.../NumeroCivicoType.xsd"/>
  <xs:element name="FormatoAcquisizione" type="FormatoAcquisizioneType"/>
  <xs:complexType name="FormatoAcquisizioneType">
    <xs:sequence>
      <xs:element name="Provincia" type="ProvinciaType">
        <xs:annotation>
          <xs:documentation> Sigla Provincia </xs:documentation>
        </xs:annotation>
      </xs:element>
      <xs:element name="Comune" type="ComuneType">
```

```
        <xs:annotation>
          <xs:documentation> Denominazione del Comune</xs:documentation>
        </xs:annotation>
      </xs:element>
      <xs:element name="Localita" type="LocalitaType" minOccurs="0">
        <xs:annotation>
          <xs:documentation> Denominazione Località. La località è obbligatoria se subcomunale</xs:documentation>
        </xs:annotation>
      </xs:element>
      <xs:element name="CAP" type="CAPType" minOccurs="0">
        <xs:annotation>
          <xs:documentation> Codice di avviamento postale, non obbligatorio</xs:documentation>
        </xs:annotation>
      </xs:element>
      <xs:element name="Denominazione" type="DenominazioneType"/>
      <xs:element name="DUG" type="DUGType"/>
      <xs:element name="NumeroCivico" type="NumeroCivicoType"/>
    </xs:sequence>
  </xs:complexType>
</xs:schema>
```

Figura 4.3: Il Formato di Acquisizione Elettronico.

Ciascuna delle componenti del dato toponomastico risulta essere di un tipo specificamente definito, sulla base delle motivazioni esposte nel corso del Paragrafo 4.2.2. Nelle Figure da 4.4 a 4.10, si riportano le definizioni esemplificate di tali tipi, definiti come XML Schema *inclusi* nel formato rappresentato in Figura 4.3; ad esempio, il tipo della Denominazione di un'area di circolazione è incluso nel formato di acquisizione in Figura 4.3 mediante la *macro* :

```
<xs:include schemaLocation="F:\Data\istat\DenominazioneType.xsd"/>
```

Le definizioni complete dei tipi delle componenti sono invece riportate in Appendice.

```
<xs:simpleType name="ProvinciaType">
      <xs:restriction base="xs:string">
        <xs:enumeration value="AG"/>
        <xs:enumeration value="AL"/>
```

```
            <!--etc...-->
        </xs:restriction>
    </xs:simpleType>
```

Figura 4.4: Esempio di definizione del tipo provincia

```
<xs:simpleType name="ComuneType">
        <xs:restriction base="xs: stringr">
          <xs:enumeration value="Agliè"/>
          <xs:enumeration value="Airasca"/>
            <!--etc...-->
        </xs:restriction>
  </xs:simpleType>
```

Figura 4.5: Esempio di definizione del tipo comune

```
<xs:simpleType name="DUGType">
            <xs:restriction base="xs:string">
              <xs:enumeration value="via"/>
              <xs:enumeration value="piazza"/>
              <!--etc...-->
            </xs:restriction>
  </xs:simpleType>
```

Figura 4.6: Esempio di definizione del tipo DUG

```
<xs:simpleType name="CAPType">
        <xs:restriction base="xs:string">
          <xs:enumeration value="00198"/>
          <xs:enumeration value="00193"/>
            <!--etc...-->
        </xs:restriction>
    </xs:simpleType>
```

Figura 4.7: Esempio di definizione del tipo CAP

```
  <xs:simpleType name="LocalitàType">
        <xs:restriction base="xs:string">
          <xs:enumeration value="AGLIE"/>
```

```
        <xs:enumeration value=" MADONNA DELLE GRAZIE "/>
          <!--etc...-->
      </xs:restriction>
.</xs:simpleType>
```

Figura 4.8: Esempio di definizione del tipo Località

```
<?xml version="1.0" encoding="UTF-8"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema">
  <xs:annotation>
    <xs:documentation>Questo schema definisce il tipo della Denominzione di un'Area di Circolazione</xs:documentation>
  </xs:annotation>
  <xs:complexType name="DenominazioneType">
    <xs:sequence>
      <xs:element name="DenominazioneSimple">
        <xs:simpleType name="DenominazioneTypeSimple">
          <xs:restriction base="xs:string">
            <xs:maxLength value="60"/>
          </xs:restriction>
        </xs:simpleType>
      </xs:element>
    </xs:sequence>
    <xs:attribute name="Lingua" type="xs:string" use="optional" default="Italiano"/>
  </xs:complexType>
</xs:schema>
```

Figura 4.9: Definizione del tipo denominazione

```
<?xml version="1.0" encoding="UTF-8"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema">
  <xs:annotation>
    <xs:documentation>Questo schema definisce la struttura dei numeri civici</xs:documentation>
  </xs:annotation>
  <xs:complexType name="NumeroCivicoType">
    <xs:sequence>
```

```
        <xs:element name="TipologiaParteNumerica" type="TipologiaParteNumericaType"/>
        <xs:element name="ParteNumerica" type="ParteNumericaType"/>
        <xs:element name="AlfaNumerico" type="AlfaNumericoType" minOccurs="0"/>
        <xs:element name="RossoNero" type="RossoNeroType" minOccurs="0"/>
        <xs:element name="CivicoAssente" minOccurs="0"/>
      </xs:sequence>
    </xs:complexType>
    <xs:simpleType name="TipologiaParteNumericaType">
      <xs:annotation>
        <xs:documentation> n=progressivo numerico e m=metri</xs:documentation>
      </xs:annotation>
      <xs:restriction base="xs:string">
        <xs:enumeration value="n"/>
        <xs:enumeration value="m"/>
      </xs:restriction>
    </xs:simpleType>
    <xs:attribute name="ParteNumericaType" type="xs:integer"/>
    <xs:attribute name="AlfaNumericoType" type="xs:string"/>
    <xs:simpleType name="RossoNeroType">
      <xs:annotation>
        <xs:documentation> r=rosso e n=nero</xs:documentation>
      </xs:annotation>
      <xs:restriction base="xs:string">
        <xs:enumeration value="r"/>
        <xs:enumeration value="n"/>
      </xs:restriction>
    </xs:simpleType>
    <xs:attribute name="CivicoAssente" use="optional" default="SNC"/>
</xs:schema>
```

Figura 4.10: Definizione del tipo numero civico

4.3 Strategia di Miglioramento della Qualità dei Dati Toponomastici basata sui Processi Interni alle Amministrazioni

Nel Capitolo 2, si è introdotto il processo generale suggerito alle amministrazioni per il miglioramento della qualità dei dati toponomastici. Il miglioramento interno alle amministrazioni risulta essere una fase di tale processo, che si colloca a valle della fase di misurazione, all'interno delle amministrazioni. In virtù dell'elevata complessità di realizzazione effettiva della fase di miglioramento interno, si rende necessario suggerire alle amministrazioni un'opportuna strategia da seguire. La strategia proposta è rappresentata in Figura 4.10. da cui si desume che il miglioramento interno è il risultato di due azioni.

Una prima azione riguarda i processi interni alle amministrazioni, che devono essere ri-progettati in maniera tale da includere controlli di qualità, basati sull'utilizzo di dizionari per le componenti dell'indirizzo e sulle regole di obbligatorietà introdotte per i formati di acquisizione (procedure generali in Figura 4.10). Inoltre, a partire da risultati di misurazione della qualità, i processi interni possono includere controlli ad-hoc, per prevenire errori nell'immissione di dati particolarmente critici.

Una seconda azione riguarda interventi di bonifica, guidati dai risultati della misurazione. In generale, l'utilizzo esclusivo di interventi di bonifica, anche se periodico, non è sufficiente a garantire un adeguato livello di qualità dei dati sul lungo termine, specialmente se l'archivio è frequentemente aggiornato. Si rende dunque necessario, complementare la bonifica dei dati con un intervento sui processi che possa, invece, garantire la qualità dei dati sul medio e lungo termine. Infatti, l'azione sui processi consente di rimuovere le cause prime degli errori di qualità, così prevenendo la possibilità di errori futuri.

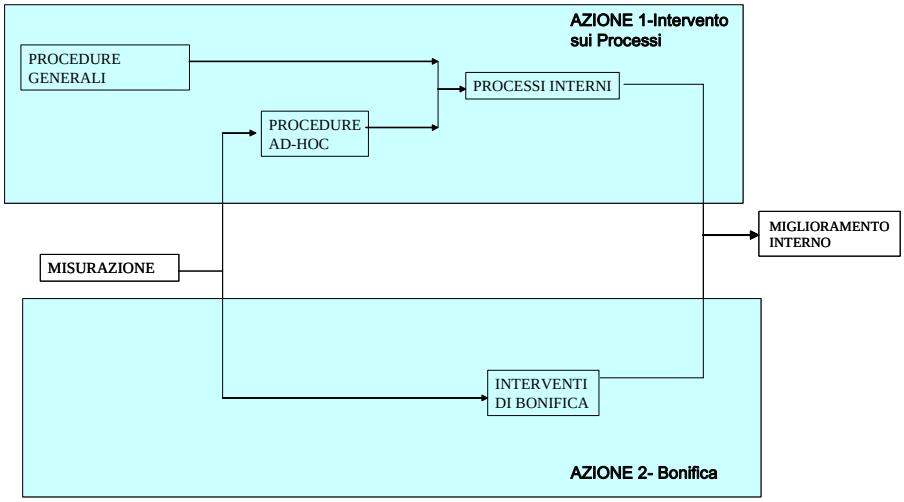


Figura 4.10: La strategia di miglioramento interno.

4.3.1 Intervento sui Processi

Si distinguono due tipologie di processi che coinvolgono la memorizzazione dei dati toponomastici all'interno di un'amministrazione:

– processi di acquisizione da utente esterno. Come discusso, l'acquisizione può avvenire mediante moduli cartacei o elettronici;

– processi di inserimento/aggiornamento nella base dati dell' amministrazione.

I processi di inserimento/aggiornamento sono collocati a valle dei processi di acquisizione. A seconda della tipologia dei processi di acquisizione, i controllo di qualità sono da effettuarsi in momenti diversi. In particolare, se il processo di acquisizione è elettronico, i controlli di qualità devono essere collocati in fase di acquisizione. Invece, qualora il processo di acquisizione sia cartaceo, i controlli di qualità devono essere realizzati nell'ambito dei processi di inserimento/aggiornamento. La Figura 4.11 riassume la collocazione dei controlli di qualità, nell'ambito dei processi di acquisizione e inserimento/aggiornamento.

I controlli di qualità si distinguono in (vedi anche Figura 4.10):

– procedure generali;

– procedure ad-hoc guidate dalla misurazione.

Nell'ambito delle procedure generali, le proposte per il miglioramento riguardano: l'*utilizzo di dizionari di riferimento* (Paragrafo 4.3.1.1) e il *controllo sull'obbligatorietà dei campi* (Paragrafo 4.3.1.2) .

Le procedure ad-hoc, che sono basate sui risultati della misurazione, sono descritte nel Paragrafo 4.3.1.3.

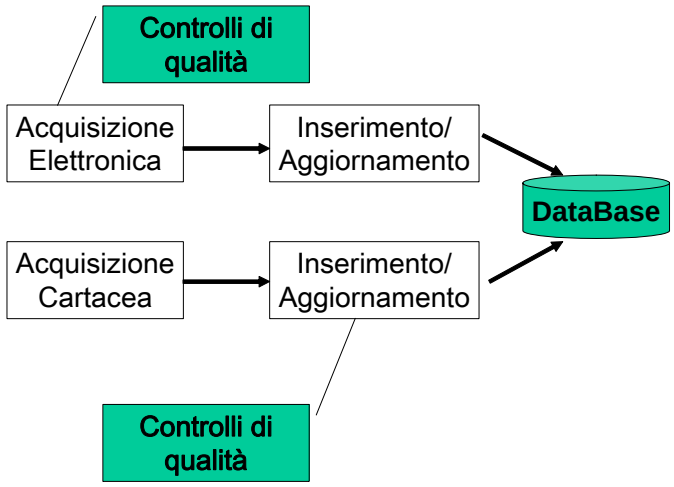


Figura 4.11: I controlli di qualità nell'ambito dei processi di acquisizione e inserimento/aggiornamento.

4.3.1.1 Procedure Generali: i Dizionari

I dizionari di riferimento sono un utile strumento da utilizzare per il controllo dei dati, da immettere nel sistema informatico di una pubblica amministrazione. In particolare, la disponibilità di un dizionario per un campo, consente il controllo dei valori ammissibili per quel campo. Il vantaggio, in termini di miglioramento della qualità dei dati, si riflette soprattutto sull'accuratezza dei valori inseriti.

Nel caso di processi di acquisizione elettronica, si suggerisce di introdurre la possibilità di selezionare il valore del campo dalla lista dei valori del dizionario relativo. Si noti come, in virtù della natura gerarchica delle singole componenti del dato toponomastico, l'insieme dei valori da selezionare ad un livello gerarchico può essere progressivamente limitato dalle scelte effettuate ai livelli gerarchici superiori. Ad esempio, una volta selezionato il valore per il campo provincia, viene automaticamente limitato il dizionario di riferimento dei comuni, ai soli comuni appartenenti alla provincia stessa.

Nel caso dei processi di inserimento/aggiornamento, il valore acquisito mediante modulo cartaceo deve essere confrontato con i valori di riferimento. Il confronto del valore acquisito con il valore di riferimento può essere effettuato mediante un applicativo software per il riconoscimento, del genere di quello proposto per la misurazione. Se il valore acquisito è riconosciuto, esso viene memorizzato nella base dati sottostante. Nel caso di valore non riconosciuto, sono possibili diverse strategie di seguito elencate:

- inserimento del valore con segnalazione di errore. Questa strategia è di immediata applicazione ma ha il chiaro svantaggio di segnalare senza correggere i dati errati nella base dati;
- non inserimento del valore errato. Questa strategia prevede l'avvio di un processo costoso, relativo ad una nuova acquisizione del dato errato, mediante riconsultazione della fonte di acquisizione. Questa strategia pur essendo costosa, garantisce la correttezza della base dati;
- inserimento e attivazione del processo di ri-acquisizione. Questa strategia prevede l'inserimento del valore, con segnalazione dell'errore e *contemporaneamente* l' attivazione della procedura di ri-acquisizione del dato. Questa strategia è quella suggerita, in quanto consente di sfruttare i vantaggi di entrambe le strategie elencate.

In Appendice, sono forniti i dizionari di riferimento per le seguenti componenti del dato toponomastico:

- Sigla provincia
- Comune
- Località
- DUG

Il dizionario relativo alle sigle delle province, contiene le sigle delle 103 province italiane attuali.

Il dizionario dei comuni, contiene la denominazione ufficiale dei comuni che l'Istat ha il compito istituzionale di rilevare. La lista dei comuni, fornita nel dizionario è quella dei comuni esistenti al 30 giugno 2003.

Il dizionario delle località comprende l'elenco di un particolare tipo di località, ovvero le località abitate (vedi Glossario), individuate dai comuni in occasione delle rilevazioni censuarie ISTAT che hanno cadenza decennale. Si noti come i comuni hanno una responsabilità stabilita dalla normativa sulle località; ciò avalla la scelta di

considerare le località fornite dai comuni. Inoltre, la scelta di fornire le località abitate è confortata dalla politica di Poste Italiane, che intende adottare il concetto di località abitata nella struttura degli indirizzi utilizzati nelle proprie basi di dati. Attualmente, non è ancora ufficializzato l'elenco delle località, individuate con il censimento del 2001, per cui l'elenco fornito con questo documento è quello relativo al Censimento del 1991.

Il dizionario delle DUG include il risultato di un complesso processo di integrazione degli elenchi delle DUG disponibili presso l'ISTAT e presso Poste Italiane. E' stata effettuata un'analisi approfondita del grado di sovrapposizione di tali elenchi, disponibili presso i due enti.

4.3.1.2 Procedure Generali: Controlli di Obbligatorietà

Nel Paragrafo 4.2 si è discussa la necessità di introdurre regole di obbligatorietà per le componenti del dato toponomastico. In particolare, un insieme di componenti è stato individuato come insieme minimo, caratterizzante l'indirizzo nel suo complesso, mediante le componenti identificate come obbligatorie.

Quando il processo di acquisizione è elettronico, devono essere effettuati dei controlli che assicurino l'immissione di un valore, per le componenti individuate come obbligatorie, nell'ambito dello specifico processo di acquisizione. L'acquisizione stessa deve essere inibita, se almeno una delle componenti obbligatorie non è specificata. Nel caso generale di acquisizione dell'indirizzo nel suo complesso, tutte le componenti: provincia, comune, denominazione area di circolazione e numero civico, devono essere specificate.

Quando il processo di acquisizione è cartaceo, l'operatore che acquisisce il modulo dalla fonte deve verificare che questo sia compilato nei campi identificati come obbligatori. Esistono, tuttavia casi in cui tale controllo non è possibile; ad esempio, nel caso di invio mediante posta ordinaria del modulo cartaceo compilato. In tali casi, se una componente specificata come obbligatoria risulta mancante, si può utilizzare un software di riconoscimento che consenta di ricostruire il valore mancante. Qualora tale software non abbia successo, si ricade in una casistica di strategie possibili del tutto simile a quella elencata nel caso di mancato riconoscimento di un valore rispetto al valore di riferimento (Paragrafo 4.3.1.1).

4.3.1.3 Procedure Ad-Hoc

Per migliorare la qualità dei dati presenti in un archivio, un passo preliminare è la misurazione della qualità. Infatti, è importante conoscere le dimensioni del problema, al fine di poter intraprendere opportune azioni risolutive. Come illustrato nel Paragrafo 2.3, il processo generale per il miglioramento della qualità dei dati toponomastici, include una fase di misurazione che le amministrazioni possono scegliere di condurre. I risultati della applicazione di tale fase di misurazione sui propri archivi interni, possono essere opportunamente utilizzati, al fine di progettare interventi *ad-hoc* sui processi interni alle amministrazioni stesse, con l'obiettivo di eliminare possibili cause di errore.

In particolare, se un'amministrazione applica la tecnica di individuazione delle aree critiche, potrebbero emergere zone degli archivi di interesse particolarmente affette da

errori sintattici, ovvero aree critiche. Per progettare interventi ad-hoc, l'amministrazione deve ricostruire i processi di acquisizione ed inserimento/ aggiornamento, che interessano i record inclusi nelle aree critiche. Se si verifica la condizione in cui un particolare processo è univocamente individuato come *causa* degli errori sintattici, presenti in un'area critica, tale processo deve essere opportunamente analizzato. L'analisi del processo deve essere mirata ad identificare la attività o le attività che sono causa di errore. A tal punto, il processo in esame deve essere riprogettato con riferimento alle specifiche attività cause di errori, al fine di eliminare o attenuare gli errori stessi. Una metodologia utile a tale fine è descritta in [SistemiInformativi2001].

Se un'amministrazione sceglie di non applicare in maniera diretta la tecnica di individuazione delle aree critiche, può comunque utilizzare i risultati nella sperimentazione descritta nel Capitolo 3. Un risultato generale mostra la dipendenza di alcuni errori sintattici sugli indirizzi dalla collocazione geografica degli indirizzi stessi. Infatti, una percentuale di errore simile si è riscontrata su tutti gli archivi analizzati nella sperimentazione condotta. Ad esempio, gli indirizzi dei comuni collocati nel Trentino Alto Adige sono caratterizzati da un elevato rischio di errore sintattico a causa del bilinguismo (italiano / tedesco). Inoltre, anche nei comuni con un numero elevato di località sub comunali e in quelli poco urbanizzati (e.g., zone rurali) il rischio di errore sintattico degli indirizzi è notevole. Un'amministrazione può, dunque, utilizzare risultati generali del tipo esemplificato, per individuare i processi causa degli errori e procedere ad una riprogettazione opportuna secondo le indicazioni sopra riportate.

4.3.2 Interventi di Bonifica

L'attività di bonifica avviene a valle della misurazione degli errori e prevede di intervenire sugli errori correggendo i dati. L'azione specifica di bonifica che si propone, prevede l'utilizzo dei risultati della fase di misurazione ed in particolare del processo di riconoscimento degli indirizzi. Come riportato nel Capitolo 3, la metodologia per la misurazione della qualità dei dati toponomastici consta, per la fase di riconoscimento degli indirizzi, di tre possibili esiti:

- indirizzo riconosciuto come corretto;
- indirizzo riconosciuto con violazione debole;
- indirizzo non riconosciuto.

Nel caso di indirizzo riconosciuto con violazione debole, il processo di bonifica deve occuparsi di sostituire l'indirizzo scorretto, memorizzato nell'archivio in esame, con il corrispondente valore ad esso associato dal software di riconoscimento.

Nel caso di indirizzo non riconosciuto, il processo di bonifica diventa più complesso. Una possibile linea di azione consiste nel consentire l'intervento da parte di un operatore umano, esperto del dominio, che possa tentare il riconoscimento *manuale* dell'indirizzo stesso. La soluzione, economicamente più svantaggiosa, prevede la nuova acquisizione del dato specifico, qualora, ad esempio, l'indirizzo localizzi uno

specifico soggetto; in tal caso, la nuova acquisizione prevede di ricontattare il soggetto stesso come fonte per l'acquisizione.

In generale, la conduzione di un processo di bonifica sull'intero archivio è un'attività economicamente onerosa. In alternativa, si suggerisce la possibilità di effettuare interventi di bonifica in maniera limitata. In particolare, si può scegliere di fissare una soglia per l'errore sintattico, al di sopra della quale intervenire con la bonifica. Le aree critiche fissano un valore di riferimento indicativo per la decisione del livello di soglia.

Si noti come, nel processo di bonifica occorre tarare opportunamente l'algoritmo o gli algoritmi di riconoscimento degli indirizzi. Inoltre, l'assenza di uno stradario nazionale, cioè di una base dati riconosciuta e gestita a livello istituzionale, contenente tutte le strade di tutti i comuni d'Italia, non assicura la completa esattezza dei risultati. Infatti, gli algoritmi utilizzati in genere prevedono il riconoscimento per similitudine, che dà la probabilità, ma non la certezza che il risultato sia esatto.

5 Miglioramento Basato sulla Cooperazione

Nell'ambito del processo proposto alle amministrazioni per il miglioramento della qualità dei dati toponomastici, la fase di miglioramento basato sulla cooperazione tra le pubbliche amministrazioni si colloca a valle della fase di misurazione, descritta nel Capitolo 3, e può avvenire in parallelo alla fase di miglioramento interno, descritta nel Capitolo 4. La fase di miglioramento basato sulla cooperazione prevede due interventi principali: *(i)* l' adozione di un formato di interscambio standard nell'ambito dei flussi inter-amministrazioni; *(ii)* una riprogettazione dei flussi informativi di scambio dei dati toponomastici al fine di garantirne l'aggiornamento.

La definizione di formati standard per i dati toponomastici è importante per controllarne e migliorarne la qualità. Nel Capitolo 4 si è descritto il formato standard di acquisizione, nel Paragrafo 5.1 del presente capitolo si illustra, invece, lo standard proposto alle amministrazioni per l'interscambio dei dati. Nel Paragrafo 5.2, si descrive uno schema generale per la riprogettazione dei flussi inter-amministrazioni.

5.1 Formato Standard di Interscambio

Il formato di interscambio proposto descrive la *rappresentazione* che ciascuna delle componenti identificate, del dato toponomastico, dovrebbe avere quando è inviata nell'ambito di flussi inter-amministrazioni.

Il formato di interscambio suggerisce anche un insieme di *metadati* che le pubbliche amministrazioni potrebbero utilizzare, al fine di migliorare la comprensione dell'informazione scambiata. Inoltre, metadati specifici sono stati introdotti per caratterizzare il livello di qualità dell'informazione scambiata.

Il formato proposto definisce le modalità di rappresentazione delle seguenti componenti del dato toponomastico:

- Provincia
- Comune
- Località
- CAP
- Denominazione Area di Circolazione
- DUG
- Numero Civico

Queste componenti sono state le medesime selezionate secondo il processo descritto nel Capitolo 4. Come evidenziato, tali componenti sono obbligatorie nella specifica di un indirizzo, con l'eccezione di CAP e località. Sono state altresì individuate alcune

condizioni particolari in cui il campo località risulta anch'esso necessario all'identificazione dell'indirizzo.

Il formato per ciascuna componente nell'interscambio risulta diverso da quello proposto per l'acquisizione. Si suppone, infatti, di adottare un interscambio automatico; in tale ipotesi, è conveniente rappresentare le componenti mediante codici ufficiali, laddove è possibile. L'utilizzo di codici consente di ottenere vantaggi sia in termini di efficienza del trasferimento dati sia in termini di rappresentazione delle componenti, mediante un riferimento ufficiale e riconosciuto. Per la rappresentazione delle componenti relative a: provincia, comune e località, si sono adottati i rispettivi codici Istat . La scelta della codifica Istat, tra le possibili codifiche disponibili, è dovuta a:

- motivazioni normative. Le codifiche Istat sono ufficiali e riconosciute dalla normativa esistente;
- motivazioni connesse alla diffusione dei codici Istat nelle basi di dati della PA. Le codifiche Istat sono facilmente accessibili per via telematica e sono già state utilizzate nell'ambito di importanti progetti per l'interscambio informativo tra le amministrazioni pubbliche, ad esempio nel progetto SAIA.

L' architettura che si considera come riferimento per l'interscambio automatico è mostrata in Figura 5.1.

Nell'ottica di abilitare la cooperazione applicativa tra le pubbliche amministrazioni, si assume un interscambio automatico dell'informazione tra le stesse. In particolare, si assume che le informazioni vengano scambiate su supporto elettronico e che tale supporto sia costituito da file XML [XML2004]. In quanto standard "de facto", per lo scambio di informazioni in contesti di cooperazione distribuita, XML non costituisce una scelta tecnologica, ma piuttosto di adesione ad un formato ampiamente adottato e non vincolante con riferimento a specifiche tecnologie di supporto.

Nel presente capitolo non si fornisce il dettaglio progettuale di un'architettura cooperativa che supporti XML come formato di interscambio, in quanto al di fuori degli obiettivi preposti. Si individua, invece, lo schema a cui i file XML di interscambio devono risultare conformi, mediante la definizione delle componenti del dato toponomastico secondo XML Schema ([XMLSchema1], [XMLSchema2]).

XML Schema è stato scelto per la rappresentazione degli schemi XML, in alternativa all'utilizzo di DTD (Document Type Definition), utilizzati in altri progetti di e-Government in Italia. Tale scelta è motivata dal fatto che, oltre ad avere le capacità espressive dei DTD, XML Schema consente la rappresentazione del tipo di dato, utile ai fini della definizione del formato di interscambio.

In Figura 5.1, tre amministrazioni si scambiano file XML che sono conformi al formato definito tramite XML Schema.

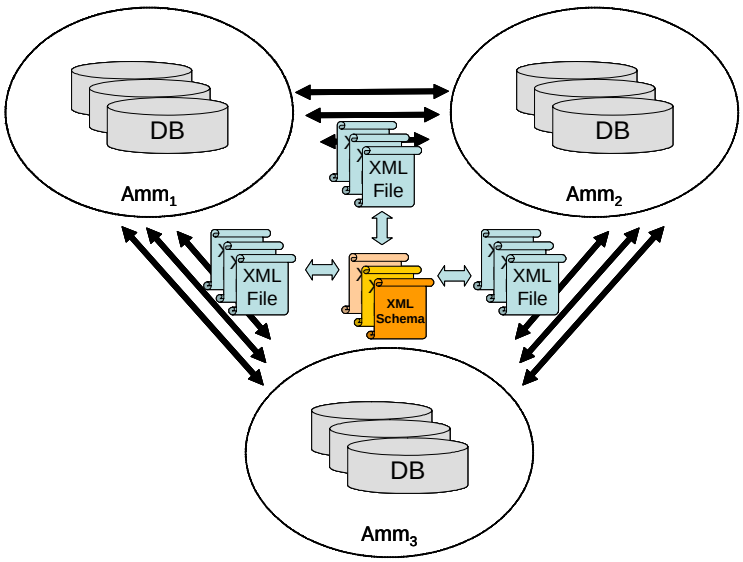


Figura 5.1: Architettura di riferimento per formato di interscambio.

Nel seguito del presente Paragrafo, per ciascuna componente dell'indirizzo verranno descritte le scelte in merito alla rappresentazione nell'interscambio. Quindi, verrà illustrata la definizione della singola componente in termini di XML Schema. Saranno illustrate inizialmente le componenti obbligatorie e poi quelle opzionali. Si noti che le considerazioni effettuate nel Capitolo 4, in merito alle caratteristiche di obbligatorietà delle componenti del dato toponomastico, continuano ad essere valide nel caso della rappresentazione di tali componenti nel formato di interscambio.

5.1.1 Componenti Obbligatorie

5.1.1.1 Provincia

Si è scelto di rappresentare il campo provincia mediante il *codice provincia Istat.* A tal fine, si è definito un elemento XML di tipo CodiceProvincia (*CodiceProvinciaType*) mediante la seguente sintassi:

```
<xs:element name="CodiceProvincia" type="CodiceProvinciaType">
    <xs:annotation>
      <xs:documentation> Codice Provincia Istat </xs:documentation>
    </xs:annotation>
</xs:element>
```

Un’annotazione riporta che le province sono rappresentate mediante il codice identificativo delle province utilizzato dall’Istat.

Il tipo CodiceProvincia è stato definito come tipo enumerato, ciascun elemento del quale rappresenta un codice Istat identificativo di una provincia, si ha, quindi, una sequenza di 3 cifre numeriche. La sintassi di definizione del tipo *CodiceProvinciaType* è riportata di seguito:

```
<xs:simpleType name="CodiceProvinciaType">
      <xs:restriction base="xs:string">
        <xs:enumeration value="001"/>
        <xs:enumeration value="002"/>
            <!--etc...-->
      </xs:restriction>
    </xs:simpleType>
```

Si noti che il tipo CodiceProvincia contiene i valori del *dizionario delle province*, riportato in Allegato. Tale dizionario mette in corrispondenza le denominazioni delle province con i codici Istat identificativi delle province. Nel dizionario delle province sono riportati i codici Istat delle province a 6 cifre che contengono anche l’identificazione della regione di appartenenza delle province nelle prime 3 cifre. Questo rende possibile effettuare controlli applicativi con riferimento ad elaborazioni che richiedono dati a livello di regione.

Il dizionario costituisce il modo in cui è possibile controllare la conformità dei valori dei codici delle province ai valori memorizzati nel dizionario delle province, quando i dati vengono codificati in XML, per essere inviati o decodificati in fase di ricezione.

5.1.1.2 Comune

Si è scelto di rappresentare il campo comune mediante il *codice comune Istat.* Si e’ definito un elemento XML di tipo CodiceComune (*ComuneType*), come riportato in seguito:

```
<xs:element name="Comune" type="CodiceComuneType">
      <xs:annotation>
        <xs:documentation> Codice Comune Istat </xs:documentation>
      </xs:annotation>
</xs:element>
```

Un’annotazione riporta che i comuni sono rappresentati mediante il codice comune dell’Istat.

Il codice comune Istat è una sequenza di 6 cifre numeriche. Le prime tre cifre rappresentano la provincia di appartenenza del comune. Le successive tre cifre rappresentano il codice identificativo del comune stesso, nell’ambito della provincia. Anche in questo caso si è scelto di definire un tipo enumerato. Ciascun elemento del tipo enumerato rappresenta un codice comune Istat ammissibile, ed è una stringa. La sintassi di definizione del tipo *CodiceComuneType* è riportata di seguito:

```
<xs:simpleType name="CodiceComuneType">
    <xs:restriction base="xs: string">
      <xs:enumeration value="001001"/>
      <xs:enumeration value="001002"/>
        <!--etc...-->
    </xs:restriction>
</xs:simpleType>
```

5.1.1.3 Denominazione Area di Circolazione

Il campo Denominazione Area di Circolazione è, insieme al numero civico, un campo del dato toponomastico, per cui non è definito un dizionario. Per il campo Denominazione si suggerisce di utilizzare il tipo semplice di XML Schema *string* ristretto ad una lunghezza massima di sessanta caratteri . La definizione del tipo *DenominazioneType* è dunque:

```
<xs:complexType name="DenominazioneType">
 <xs:sequence>
 <xs:element name="DenominazioneSimple">
  <xs:simpleType name="DenominazioneTypeSimple">
    <xs:restriction base="xs:string">
      <xs:maxLength value="60"/>
    </xs:restriction>
  </xs:simpleType>
</xs:element>
  </xs:sequence>
  <xs:attribute name="Lingua" default="Italiano" type="xs:string" use="optional"/>
  </xs:complexType>
```

```
</xs:schema>
```

Si noti la presenza di un attributo *Lingua* nella specifica del campo, il cui significato sarà chiarito nella sezione dedicata ai metadati.

5.1.1.4 DUG

Il campo DUG è definito mediante il tipo DUG (*DugType*) che è un'enumerazione dei valori ammissibili per una DUG e riportati nel *dizionario delle DUG*, in Allegato. La definizione del tipo DUG è di seguito riportata:

```
<xs:simpleType name="DUGType">
        <xs:restriction base="xs:string">
          <xs:enumeration value="via"/>
          <xs:enumeration value="piazza"/>
          <!--etc...-->
        </xs:restriction>
 </xs:simpleType>
```

5.1.1.5 Numero Civico

Analogamente a quanto discusso nel Capitolo 4 per il formato di acquisizione, anche nel formato di interscambio il campo Numero Civico è definito come una sequenza di cinque campi. Inoltre, le regole di obbligatorietà tra tali campi, riportate nella tabella 4.3 del Capitolo 4, continuano ad essere valide anche per il formato di interscambio. I singoli campi e la loro rappresentazione sono di seguito elencati:

- *Tipologia Parte Numerica*. Consente di discriminare i casi in cui la parte numerica rappresenta un progressivo numerico, ovvero un'indicazione chilometrica espressa in metri. Un'annotazione è esplicitamente prevista allo scopo di illustrare il ruolo ed i possibili valori di questo campo. Il tipo del campo Tipologia Parte Numerica (TipologiaParteNumericaType) è definito come un tipo enumerato con due valori ammissibili: *(i)* n, indicante che la parte numerica rappresenta un progressivo numerico; *(ii)* m, indicante che la parte numerica rappresenta un'indicazione chilometrica espressa in metri.
- *Parte Numerica*. Contiene la parte numerica del numero civico. Il tipo della Parte Numerica (ParteNumericaType) è il tipo semplice *integer*.
- *Parte Alfanumerica*. Contiene la parte alfanumerica del numero civico. Il tipo della Parte Alfanumerica (ParteNumericaType) e' definito come tipo semplice di XML Schema *string*.

– *RossoNero*. Considera la possibilità di avere una classificazione dei civici sulla base dei due possibili valori rosso e nero. Il tipo del campo RossoNero (RossoNeroType) è definito come un tipo enumerato con due valori ammissibili: *(i)* `r`, indicante rosso; *(ii)* n, indicante nero. Un'annotazione è esplicitamente prevista allo scopo di illustrare il ruolo ed i possibili valori di questo campo.

– *CivicoAssente*. Evidenzia i casi in cui non esiste un numero civico. Qualora il civico non esista, l'esplicita dichiarazione di tale status, mediante la presenza di tale campo che ha come valore di default *snc*, i.e. senza numero civico.

La sintassi di definizione del tipo NumeroCivicoType è di seguito riportata.

```
<?xml version="1.0" encoding="UTF-8"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema">
  <xs:annotation>
    <xs:documentation>Questo schema definisce la struttura dei numeri civici</xs:documentation>
  </xs:annotation>
<xs:complexType name="NumeroCivicoType">
    <xs:sequence>
      <xs:element name="TipologiaParteNumerica" type="TipologiaParteNumericaType"/>
      <xs:element name="ParteNumerica" type="ParteNumericaType"/>
      <xs:element name="AlfaNumerico" type="AlfaNumericoType" minOccurs="0"/>
      <xs:element name="RossoNero" type="RossoNeroType" minOccurs="0"/>
      <xs:element name="CivicoAssente" minOccurs="0"/>
    </xs:sequence>
  </xs:complexType>
  <xs:simpleType name="TipologiaParteNumericaType">
    <xs:annotation>
      <xs:documentation> n=progressivo numerico e m=metri</xs:documentation>
    </xs:annotation>
    <xs:restriction base="xs:string">
      <xs:enumeration value="n"/>
      <xs:enumeration value="m"/>
    </xs:restriction>
  </xs:simpleType>
```

```
  <xs:attribute  name="ParteNumericaType" type="xs:integer"/>
   <xs:attribute name="AlfaNumericoType" type="xs:string"/>
      <xs:simpleType name="RossoNeroType">
    <xs:annotation>
      <xs:documentation> r=rosso e n=nero</xs:documentation>
    </xs:annotation>
    <xs:restriction base="xs:string">
      <xs:enumeration value="r"/>
      <xs:enumeration value="n"/>
    </xs:restriction>
  </xs:simpleType>
  <xs:attribute name="CivicoAssente" use="optional" default="SNC"/>
</xs:schema>
```

5.1.2 Componenti Non Obbligatorie

5.1.2.1 Codice di Avviamento Postale

Il tipo di CAP (*CAPType*) è definito come enumerazione dei valori dei CAP corrispondenti all'elenco attuale dei CAP italiani (vedi Appendice). La definizione del tipo *CAPType* è di seguito riportata:

```
<xs:simpleType name="CAPType">
      <xs:restriction base="xs:string">
        <xs:enumeration value="00198"/>
        <xs:enumeration value="00193"/>
          <!--etc...-->
      </xs:restriction>
    </xs:simpleType>
```

5.1.2.2 Località

Si è scelto di rappresentare il campo località mediante il *codice località Istat.* Nella definizione dell'elemento località, un'annotazione esplicita il caso di obbligatorietà

della specifica località come componente del dato toponomastico[6]. La sintassi di definizione dell'elemento CodiceLocalità è di seguito riportata:

```
<xs:element name="CodiceLocalita" type="xs: CodiceLocalitaType" minOccurs="0">
    <xs:annotation>
      <xs:documentation>Codice Località Istat. La località è obbligatoria se subcomunale<
        /xs:documentation>
    </xs:annotation>
</xs:element>
```

Il codice località Istat è una sequenza di 10 cifre numeriche. Le prime tre cifre rappresentano la provincia di appartenenza della località. Le successive tre cifre rappresentano il codice identificativo del comune di appartenenza. Infine, le ultime quattro cifre rappresentano la località stessa. Anche in questo caso si è scelto di definire un tipo enumerato. Ciascun elemento del tipo enumerato rappresenta codice località Istat ammissibile, ed è una stringa.

Il tipo Località (*CodiceLocalitàType*) è definito come segue:

```
<xs:simpleType name="CodiceLocalitàType">
    <xs:restriction base="xs:string">
      <xs:enumeration value=" 0010011001"/>
      <xs:enumeration value=" 0010011002 "/>
        <!--etc...-->
    </xs:restriction>
.</xs:simpleType>
```

5.1.3 I Metadati nel Formato di Interscambio

Il ruolo generale dei *metadati* è quello di fornire informazioni sui dati applicativi. Come i dati cui si riferiscono, anche i metadati possono essere non strutturati, semi-strutturati o totalmente strutturati, a seconda della presenza o meno di una informazione intenzionale, fornita da uno schema. Un'informazione testuale, di tipo

[6] Si ricorda che esiste un caso in cui il campo località è obbligatorio. Tale caso corrisponde alle località subcomunali che consentono di individuare delle aree di circolazione, i.e. possono esistere aree di circolazione con la stessa denominazione nell'ambito dello stesso comune, ma appartenenti a due località subcomunali diverse.

“libero”, costituisce un esempio di informazione non strutturata, in virtù di una totale assenza di schema. I file XML possono essere di tipo semi-strutturato, qualora agli elementi con schema si alternino ad elementi senza schema. Infine, i dati di un database relazionale costituiscono un esempio di informazione strutturata, perché forniti di schema.

Pur utilizzando XML come formato di interscambio, le informazioni toponomastiche che le amministrazioni si scambiano nell’ambito dei propri flussi informativi sono informazioni di tipo strutturato. Si ritiene opportuno adottare per i metadati relativi ai dati toponomastici, la struttura più flessibile dei dati semistrutturati. Le motivazioni di tale scelta sono le seguenti:

– l’utilizzo di XML come formato di interscambio per i dati toponomastici, consente di sfruttare, con riferimento ai metadati, la proprietà di XML di essere “human readable”, ovvero comprensibile all’utente umano. Si ha dunque la possibilità di descrivere il significato di un campo mediante testo libero, assicurando così una buona comprensibilità del significato dei campi che sono scambiati. Questa possibilità può essere molto utile al fine di risolvere i problemi di eterogeneità semantica che caratterizzano i dati posseduti dalle amministrazioni pubbliche. Ad esempio, aggiungendo una descrizione testuale del significato di DUG, si facilita la comprensione, da parte di un’amministrazione che riceve un file XML, che contiene tale campo, del fatto che una DUG specifica la tipologia di un’area di circolazione (e.g., via, piazza, etc.);

– i dati semistrutturati consentono di attribuire uno schema all’informazione, laddove necessario. Ad esempio, nel trasferire un insieme di metadati riguardanti la *qualità* dell’informazione scambiata, si può decidere esattamente quali dimensioni di qualità devono caratterizzare i dati, mediante l’introduzione di uno schema per esse. L’immediato vantaggio di tale strutturazione è la garanzia, per un opportuno insieme di metadati, della *non-ambiguità* nell’interpretazione.

Sulla base di tali considerazioni l’insieme dei metadati previsti per il formato di interscambio sono:

– *Annotations.* Sono informazioni non strutturate descrittive di alcune componenti del dato toponomastico. Nel Paragrafo precedente alcune *annotations* sono state descritte con riferimento in particolare alle informazioni non autoesplicative: codice comune Istat, valori numerico e chilometrico per il numero civico etc. Alcune *annotations* di carattere più generale sono usate per documentare il contenuto di diversi XML Schema (si vedano ad esempio, in seguito, i file descrittivi del formato di interscambio, annotati con un’informazione che ne descrive il contenuto).

– *Attributo lingua*. Poiché, il campo Denominazione può contenere valori di una lingua diversa dall’Italiano, il campo Denominazione prevede un attributo Lingua che fornisce la meta informazione sulla lingua, dei valori di quel campo. La specifica del tipo della Denominazione include dunque:

```
<xs:attribute name="Lingua" default="Italiano" type="xs:string" use="optional">
</xs:attribute>
```

– *Glossario*. Un metadato indicante l'URL per un glossario di riferimento delle componenti del dato toponomastico completa l'insieme dei metadati di corredo agli scambi di dati toponomastici. La specifica dell'URL può avvenire, ad esempio, come segue:

```
<xs:element name="GlossarioTermini" type="anyURI" fixed="http://www.Istat.it/AIPA-Istat" />
```

– *Metadati di Qualità* .Oltre ai metadati di tipo specifico sopra descritti, il documento di interscambio dell'informazione toponomastica dovrà contenere una *sezione metadati* di qualità, all'inizio del documento. La sezione metadati di qualità contiene le seguenti meta-informazioni. Ciascuna delle componenti del dato toponomastico, diversa da un codice[7], ha associato un insieme di metadati che ne caratterizzano il livello di qualità. In particolare, i metadati di qualità proposti sono*: (i) un time-stamp di last updat,* che consiste nella data ultimo aggiornamento di tutti i valori della specifica componente; *(ii)* un valore di accuratezza; *(iii)* un valore di completezza e *(iv)* un valore di consistenza. Tutti i metadati di qualità devono essere specificati nella sezione metadati come segue:

```
<xs:complexType name="QualityType">
    <xs:sequence>
      <xs:element name="lastUpdate" type="date" minOccurs="0"/>
      <xs:element name="accuratezza" minOccurs="0"/>
      <xs:element name="completezza" minOccurs="0"/>
      <xs:element name="consistenza" minOccurs="0"/>
    </xs:sequence>
  </xs:complexType>
  <xs:element name="Metadati" type="MetadatiType">
    <xs:complexType name="MetadatiType">
      <xs:sequence>
        <xs:element name="DenominazioneQualità" type="QualityType"/>
```

[7] Per i codici, si suppone venga fatto a-priori un controllo di qualità con conseguente allineamento. Anche i CAP sono considerati come codici, e pertanto non ne viene esplicitamente specificata la qualità.

```
        <xs:element name="DUGQualità"/>
        <xs:element name="NumeroCivicoQualità"/>
        <xs:element name="GlossarioTermini" type="anyURI"
fixed="http://www.Istat.it/AIPA-Istat"/>
      </xs:sequence>
    </xs:complexType>
  </xs:element>
```

Si noti la cardinalità minima 0 per i metadati di qualità, indicante l'opzionalità di ciascun campo. Si noti inoltre, come l'unico campo per il quale risulta specificato il tipo è il campo di last update, il cui tipo deve essere una data. Questa scelta è voluta al fine di consentire anche solo una descrizione testuale della qualità dei dati scambiati, almeno fino al momento in cui uno standard specifico per i metadati di qualità non sarà disponibile. Tuttavia, si suggerisce di denotare la qualità dei dati toponomastici mediante gli indici descritti nel Capitolo 3, qualora sia stata opportunamente condotta la fase di misurazione della qualità.

5.1.4 XML Schema Complessivo del Formato di Interscambio

Si riporta nel seguito lo schema complessivo del formato di interscambio dei dati toponomastici.

```
<?xml version="1.0" encoding="UTF-8"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema">
  <xs:annotation>
    <xs:documentation>Questo file contiene la definizione del formato di
interscambio di ciascuna componente del dato toponomastico</xs:documentation>
  </xs:annotation>
  <xs:include schemaLocation="http://www.istat.it/.../ListaCodiciProvince.xsd"/>
  <xs:include schemaLocation=" http://www.istat.it/.../ListaCodiciComuni.xsd"/>
  <xs:include schemaLocation=" http://www.istat.it/.../ListaCodiciLocalita.xsd"/>
  <xs:include schemaLocation=" http://www.istat.it/.../ListaCAP.xsd"/>
  <xs:include schemaLocation=" http://www.istat.it/.../DenominazioneType.xsd"/>
  <xs:include schemaLocation=" http://www.istat.it/.../ListaDUG.xsd"/>
  <xs:include schemaLocation http://www.istat.it/.../NumeroCivicoType.xsd"/>
  <xs:element name="FormatoInterscambio" type="FormatoInterscambioType"/>
  <xs:complexType name="FormatoInterscambioType">
    <xs:sequence>
```

```
      <xs:element name="CodiceProvincia" type="CodiceProvinciaType">
        <xs:annotation>
          <xs:documentation> Codice Provincia Istat</xs:documentation>
        </xs:annotation>
      </xs:element>
      <xs:element name="CodiceComune" type="CodiceComuneType">
        <xs:annotation>
          <xs:documentation> Codice Comune Istat</xs:documentation>
        </xs:annotation>
      </xs:element>
      <xs:element name="CodiceLocalita" type="CodiceLocalitaType" minOccurs="0">
        <xs:annotation>
          <xs:documentation> Codice Localita Istat. La località è obbligatoria se subcomunale</xs:documentation>
        </xs:annotation>
      </xs:element>
      <xs:element name="CAP" type="CAPType" minOccurs="0">
        <xs:annotation>
          <xs:documentation> Codice di avviamento postale, non obbligatorio</xs:documentation>
        </xs:annotation>
      </xs:element>
      <xs:element name="Denominazione" type="DenominazioneType"/>
      <xs:element name="DUG" type="DUGType"/>
      <xs:element name="NumeroCivico" type="NumeroCivicoType"/>
    </xs:sequence>
  </xs:complexType>
</xs:schema>
```

5.1.5 XML Schema Complessivo del Formato di Interscambio con Metadati

Si riporta nel seguito lo schema complessivo del formato di interscambio dei dati toponomastici, insieme alla specifica dei metadati.

```
<?xml version="1.0" encoding="UTF-8"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema">
  <xs:annotation>
    <xs:documentation>Questo file contiene la definizione del formato di interscambio di ciascuna componente del dato toponomastico e l' insieme di metadati associato </xs:documentation>
  </xs:annotation>
  <xs:include schemaLocation=" http://www.istat.it/.../ListaCodiciProvince.xsd"/>
  <xs:include schemaLocation=" http://www.istat.it/.../ListaCodiciComuni.xsd"/>
  <xs:include schemaLocation=" http://www.istat.it/.../ListaCodiciLocalita.xsd"/>
  <xs:include schemaLocation=" http://www.istat.it/.../ListaCAP.xsd"/>
  <xs:include schemaLocation=" http://www.istat.it/.../DenominazioneType.xsd"/>
  <xs:include schemaLocation=" http://www.istat.it/.../ListaDUG.xsd"/>
  <xs:include schemaLocation=" http://www.istat.it/.../NumeroCivicoType.xsd"/>
   <xs:include schemaLocation=" http://www.istat.it/.../FormatoInterscambio.xsd"/>
<xs:complexType name="QualityType">
    <xs:sequence>
      <xs:element name="lastUpdate" type="date" minOccurs="0"/>
      <xs:element name="accuratezza" minOccurs="0"/>
      <xs:element name="completezza" minOccurs="0"/>
      <xs:element name="consistenza" minOccurs="0"/>
    </xs:sequence>
  </xs:complexType>
  <xs:element name="Metadati" type="MetadatiType">
    <xs:complexType name="MetadatiType">
      <xs:sequence>
        <xs:element name="DenominazioneQualità" type="QualityType"/>
        <xs:element name="DUGQualità" type="QualityType"/>
        <xs:element name="NumeroCivicoQualità" type="QualityType"/>
        <xs:element name="GlossarioTermini" type="anyURI" fixed="http://www.Istat.it/AIPA-Istat"/>
      </xs:sequence>
    </xs:complexType>
```

```
  </xs:element>
  <xs:element name="FormatoInterscambio" type="FormatoInterscambioType"/>
</xs:schema>
```

5.2 Analisi e Progettazione dei Flussi Informativi Connessi a Dati Toponomastici

Migliorare la qualità dei dati è una problematica complessa. Non è infatti sufficiente che un'amministrazione si limiti a *misurare* la qualità delle proprie basi di dati. La misurazione è solo un passo propedeutico ad interventi di *miglioramento* della qualità dei dati, che in maniera più sostanziale e definitiva consentono la gestione della qualità dei dati nel tempo. Nel Capitolo 4 si è descritto come effettuare il miglioramento della qualità dei dati toponomastici all'interno delle amministrazioni. Nel presente Paragrafo si affronta, invece, il problema di migliorare la qualità, considerando la cooperazione tra le amministrazioni come elemento abilitante.

In particolare, nel considerare il problema del miglioramento della qualità dei dati, è necessario osservare come, con riferimento ad una singola amministrazione, i dati fluiscono attraverso differenti strutture e processi, sia interni che esterni all'amministrazione stessa. In maniera più specifica, si possono distinguere flussi informativi interni ad un'amministrazione, *flussi intra-amministrazione* e flussi informativi tra amministrazioni diverse, *flussi inter-amministrazioni*.

Nel presente Paragrafo si descrive una possibile strategia di miglioramento della qualità dei dati toponomastici basata sull'analisi e il ridisegno dei flussi *inter-amministrazioni*.

La strategia proposta si articola in due fasi sequenziali:

- la prima fase è la *Fase di Analisi* e consiste nell'identificare e classificare: *(i)* le tipologie di informazioni toponomastiche oggetto d'indagine, ovvero scambiate nell'ambito di flussi inter-amministrazioni; *(ii)* i soggetti coinvolti nello scambio di tali informazioni e *(iii)* gli eventi connessi ad operazioni di creazione ed aggiornamento delle informazioni toponomastiche considerate;
- la seconda fase è la *Fase di Progettazione dei Flussi* e prevede la progettazione dei flussi informativi connessi a dati toponomastici. Laddove attuati, tali flussi comportano un miglioramento della qualità dei dati. Come formalismo per la descrizione dei flussi progettati, si è scelto di utilizzare il linguaggio Unified Modeling Language (UML) [UML2004], principalmente perché, ormai linguaggio standard "de facto" nella modellazione di sistemi.

5.2.1 Fase di Analisi: Classificazione Dati

I dati toponomastici sono in generale connessi all'identificazione di un luogo fisico sul territorio. Ricordiamo che i dati toponomastici sono classificabili in due categorie: *dati topomastici puri o indirizzi* e *dati toponomastici per la localizzazione*.

I dati toponomastici puri sono dati memorizzati in maniera indipendente da dati di altro tipo e al solo scopo di individuare un luogo fisico sul territorio. Ad esempio, le informazioni memorizzate dai catasti comunali costituiscono dati toponomastici puri.

I dati toponomastici per la localizzazione sono invece dati logicamente memorizzati insieme ad informazioni identificative di soggetti, ed il cui scopo è quello di "localizzare" i soggetti stessi. I dati toponomastici per la localizzazione si distinguono in base alle tipologie di soggetti che sono localizzati sul territorio. In particolare, si distinguono dati toponomastici per la *localizzazione di soggetti fisici* e dati toponomastici per la *localizzazione di soggetti giuridici*. Un esempio di dato toponomastico per la localizzazione di soggetti fisici è costituito dall' *indirizzo di residenza*.

Nell'ambito dello studio condotto sui flussi informativi tra le pubbliche amministrazioni ai fini di migliorare la qualità dei dati toponomastici, il dato toponomastico puro riveste un ruolo fondamentale.

Nel seguito si prenderanno in esame tutti i flussi connessi ad eventi di creazione o modifica dei soli dati toponomastici puri. Tuttavia, essendo i dati toponomastici di localizzazione, ovviamente collegati agli eventi connessi ai dati toponomastici puri, si considereranno, altresì, gli *effetti che gli eventi sui dati toponomastici puri avranno sui dati toponomastici di localizzazione.* Il motivo di tale scelta risiede nel fatto che nell'ambito del progetto SAIA [SAIA], in relazione ai soggetti fisici e del progetto RAE [RAE], in relazione ai soggetti giuridici, è già avvenuta la definizione dei flussi informativi connessi ad eventi caratterizzanti l'informazione toponomastica di localizzazione. Ad esempio, sono stati definiti i flussi connessi all'evento di variazione di un indirizzo di residenza (dato toponomastico di localizzazione di soggetti fisici).

In Figura 5.2, è riportata la struttura generale del dato toponomastico. Per il dato toponomastico puro è riportata la suddivisione nelle componenti obbligatorie, identificate nei capitoli precedenti. Si noti che il simbolismo UML indica che il dato toponomastico puro è un'aggregazione di componenti, costituite da: comune, provincia e area di circolazione; l' area di circolazione è a sua volta un'aggregazione della denominazione dell'area e del numero civico.

.

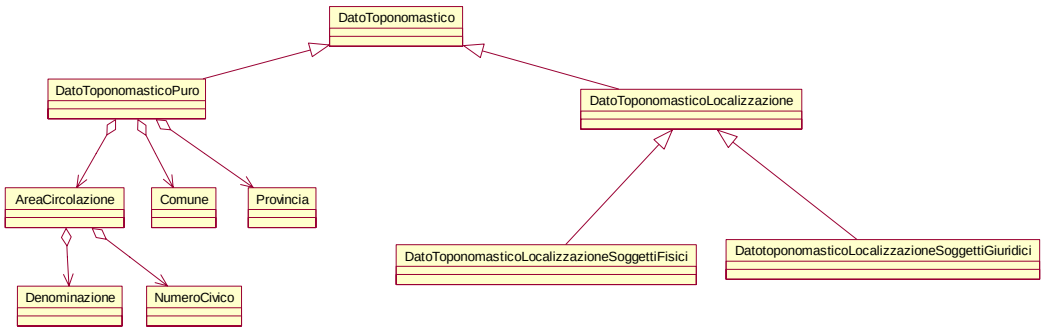


Figura 5.2: Composizione del dato toponomastico.

5.2.2 Fase di Analisi: Classificazione Soggetti

I soggetti coinvolti nei flussi informativi connessi ai dati toponomastici sono classificati in base alla tipologia di informazione che memorizzano. Ne consegue, quindi, che la classificazione dei soggetti ricalca le categorie individuate per i dati toponomastici; in particolare si distinguono:

– enti della Pubblica Amministrazione che memorizzano informazioni toponomastiche per la localizzazione dei soggetti fisici –*EntiPASoggettiFisici.* Un esempio è fornito dal Ministero dell'Economia e delle Finanze, che memorizza dati toponomastici per la localizzazione di soggetti fisici nell'Anagrafe Tributaria;

– enti della Pubblica Amministrazione che memorizzano informazioni toponomastiche per la localizzazione dei soggetti giuridici - *EntiPASoggettiGiuridici*. Un esempio è fornito dalle Camere di Commercio, che hanno basi di dati omogenee distribuite sul territorio che includono informazione toponomastica per la localizzazione di soggetti giuridici;

– enti della Pubblica Amministrazione che memorizzano informazioni toponomastica pura-*EntiPAToponomastici.* Un esempio è fornito dal Ministero dell'Economia e delle Finanze che memorizza informazione toponomastica pura nel Catasto.

In Figura 5.3 si riporta una tabella che riassume la classificazione dei soggetti rispetto alle tipologie di informazioni toponomastiche.

Tipologia Informazione/ TipologiaEnte	**Pura**	**Localizz. Sogg. Fisici**	**Localizz. Sogg. Giur.**
EntiPAToponomastici	X		
EntiPASoggFisici		X	
EntiPASoggGiuridici			X

Figura 5.3: Classificazione dei soggetti rispetto alle tipologie di informazioni toponomastiche trattate.

Si noti che la classificazione proposta non prevede classi disgiunte. Infatti, il Ministero dell'Economia e delle Finanze è classificato sia come EntePAToponomastico che come EntePASoggettiFisici, con riferimento ad archivi differenti posseduti dall'Ente.

5.2.2.1 Gli Indici Nazionali e l'Archivio Nazionale dei Dati Toponomastici

Oltre alle categorie di soggetti elencati, sono coinvolti nei flussi informativi connessi ai dati toponomastici due indici nazionali realizzati nell'ambito dei progetti SAIA [SAIA] e RAE [RAE].

In particolare si considerano gli indici nazionali:

- Indice Nazionale dei Soggetti Fisici-SAIA
- Indice Nazionale dei Soggetti Giuridici-RAE

Inoltre, un punto cardine della progettazione dei flussi, che verrà descritta nel Paragrafo 5.2, è la proposta di un *archivio nazionale dei dati toponomastici*. L'esigenza di avere un elemento logicamente centralizzato per la memorizzazione dell'informazione toponomastica pura è già esplicitata nella normativa vigente. In particolare, l'articolo 226 del Decreto Legislativo del 30 aprile 1992, n.285 (Nuovo Codice della Strada) prevede che, presso il Ministero delle Infrastrutture e dei Trasporti, sia *"istituito l'archivio nazionale delle strade, che comprende tutte le strade distinte per categorie, come indicato nell'art. 2"*. Come riportato nell'Appendice relativa alla normativa vigente, la realizzazione di questo archivio non è stata ancora ultimata.

Nella progettazione dei flussi si assumerà la presenza di un archivio nazionale dei dati toponomastici, descrivendo in particolare i flussi di alimentazione di tale archivio. La realizzazione effettiva dell'archivio potrà beneficiare di tale progettazione e dovrà inoltre confrontarsi con i risultati raggiunti dal Ministero delle Infrastrutture e dei Trasporti per la costituzione dell'archivio nazionale delle strade previsto dalla citata normativa vigente.

5.2.2.2 I Data Steward dei Dati Toponomastici

Nell'ambito dei soggetti coinvolti nei flussi connessi allo scambio di dati toponomastici, è utile identificare una figura che la letteratura sulla qualità dei dati denomina *Data Steward* [Redman96].

Il *data steward* è il soggetto cui compete la responsabilità di una specifica tipologia di informazione, secondo quanto stabilito o da una norma legislativa o in generale da altre forme di accordo tra soggetti che condividono e scambiano informazione in contesti di cooperazione.

Da un esame della normativa vigente, in materia di dati toponomastici, è stato possibile identificare il *data steward* per ciascuna delle componenti dell'indirizzo, riportate in Figura 5.2. Inoltre, anche per alcune informazioni toponomastiche per la localizzazione di soggetti fisici e giuridici sono stati individuati i *data steward*, sebbene non per singole componenti, ma per l'intera informazione.

L'individuazione dei *data steward* delle componenti del dato toponomastico puro è un passo fondamentale per la progettazione dei flussi. Infatti, come si illustrerà nel Paragrafo 5.2, il *data steward* di una tipologia di dato è il soggetto che ha la responsabilità di attivare i flussi di creazione e modifica connessi ad eventi associati alla tipologia di dato stessa.

I *data steward* individuati sono di seguito elencati:

- per la componente **Comune** del dato toponomastico puro, il *data steward* è rappresentato dalla **Regione.** Pertanto, un evento del tipo *variazione di denominazione di un comune* comporta che sia la regione, in cui il comune si colloca, ad attivare un flusso di aggiornamento del mutato valore della denominazione del comune in tutte le Pubbliche Amministrazioni interessate dall'evento;
- per la componente **Provincia** del dato toponomastico puro, il *data steward* è rappresentato: (i) dal **Ministero degli Interni** per *regioni a statuto ordinario* e, (ii) dalla **Regione** per le *regioni a statuto speciale, il cui statuto ha valore costituzionale*;
- per le componenti **Area di Circolazione** e **Numero Civico** del dato toponomastico puro, il data steward è rappresentato dall'Ufficio Toponomastico dei **Comuni**. Il regolamento anagrafico, D.P.R. 30 maggio 1989 n. 223 è particolarmente preciso nell'identificazione della responsabilità della gestione di tali tipologie di dato, individuando nell'ufficiale dell'anagrafe il soggetto fisico avente tale responsabilità. Inoltre, viene anche precisato che laddove esistano uffici toponomastici all'interno del comune, diversi da quello anagrafico, tali uffici devono comunicare a quest'ultimo *"le disposizioni ed i provvedimenti, da essi presi, concernenti l'onomastica delle aree di circolazione e la numerazione civica"* (art.44 del regolamento anagrafico di cui sopra) ;
- nell'ambito dei dati toponomastici per la localizzazione di soggetti fisici, si è individuato il *data steward* per l'**Indirizzo di Residenza** che è rappresentato dall'Ufficio Anagrafico dei **Comuni**;
- nell'ambito dei dati toponomastici per la localizzazione di soggetti giuridici, si è individuato il *data steward* per la **Sede Legale,** che è rappresentato dalle **Camere di Commercio**.

5.2.3 Fase di Analisi: Classificazione degli Eventi

Gli eventi connessi ai dati toponomastici sono classificati rispetto ai flussi di scambio dei dati toponomastici stessi. In particolare, in base a tale classificazione gli eventi si distinguono in:

- **eventi di origine**, che determinano l'inizio del flusso. Tali eventi sono connessi alla creazione e alla variazione delle *componenti del dato toponomastico puro*;
- **eventi di diffusione,** che si originano a valle degli eventi di origine e hanno lo scopo di aggiornare gli archivi interessati dall'occorrenza degli eventi di origine. Gli eventi di diffusione sono connessi ad aggiornamenti, sia di *componenti del dato toponomastico puro* che di *dati toponomastici per la localizzazione*.

Nella Figura 5.4, sono riportate le relazioni che sussistono tra gli eventi delle due tipologie elencate e i dati toponomastici puri e di localizzazione. Si noti, in particolare, come gli eventi di diffusione possono essere in generale connessi sia a dati toponomastici puri che di localizzazione; un evento specifico però è connesso in modo esclusivo ad una delle due tipologie (OR esclusivo).

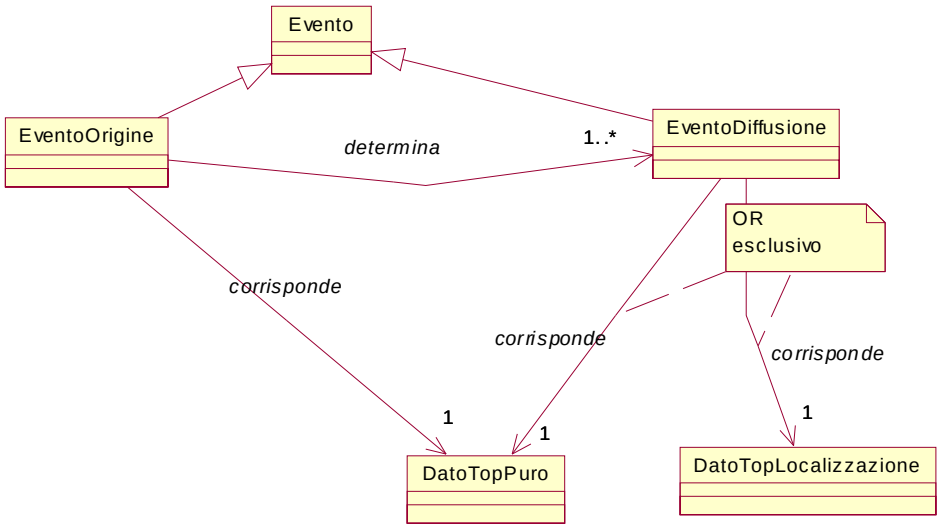
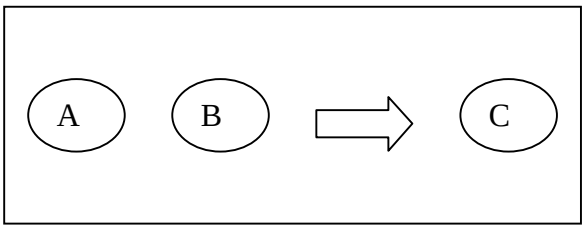


Figura 5.4: Associazioni tra tipologie di eventi e tipologie di dati toponomastici.

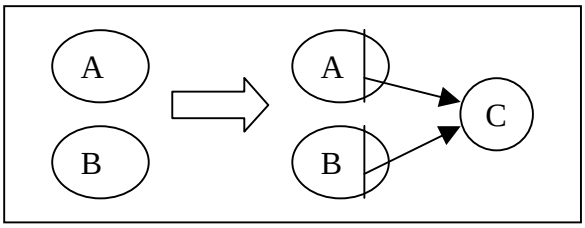
A differenza degli eventi di diffusione, gli eventi di origine possono essere distinti in varie tipologie. In particolare, gli eventi di origine sono classificabili nelle cinque seguenti tipologie:

- **Cambio di denominazione.** Questo evento è connesso ad un cambiamento di denominazione di un comune, una provincia o un'area di circolazione, laddove tale cambiamento non comporti variazioni territoriali, ovvero acquisizioni e/o cessioni di parti di territorio. Nel caso di numero civico, il cambio di denominazione si riduce alla semplice variazione del numero civico.
- **Nascita.** Questo evento può interessare un comune, una provincia o un'area di circolazione. La nascita puo' avvenire: a.1) "ex novo", ad esempio nel caso in cui ci sia una donazione allo Stato di un territorio privato; a.2) per fusione di N comuni o province o aree di circolazione (con scomparsa degli N); a.3) per acquisizione di territori da N comuni o province o aree di circolazione (che non scompaiono ma cedono una parte del loro territorio). I casi a.2) e a.3) sono rappresentati in Figura 5.5 (con N=2). L'evento nascita può anche interessare un numero civico, nel qual caso consiste semplicemente nella assegnazione ex novo del numero civico ad una nuova unità ecografica semplice (i.e. abitazione, esercizio, ufficio, etc.).
- **Scomparsa.** Questo evento può interessare un comune, una provincia o un'area di circolazione. Un comune, una provincia, o un'area di circolazione può scomparire se acquisito completamente da uno o piu' comuni o province o aree di circolazione. L'evento scomparsa può anche interessare un numero civico, nel qual caso consiste semplicemente nella cancellazione, del numero civico, da una

unità ecografica semplice non più esistente.



a.2)



a.3)

Figura 5.5: Esempi di tipologie di eventi di nascita.

- **Acquisizione di territorio.** Questo evento può interessare un comune, una provincia o un'area di circolazione. Un comune può acquisire un insieme di aree di circolazione (se l'insieme delle aree di circolazione costituisce un comune determina l'evento di scomparsa del comune stesso); una provincia può acquisire uno o più comuni (se l'insieme dei comuni costituisce una provincia determina l'evento di scomparsa della provincia stessa), e, infine, un'area di circolazione può acquisire una o più aree di circolazione (causandone la scomparsa) o porzioni di un'area di circolazione.
- **Cessione di territorio.** Questo evento può interessare un comune, una provincia o un'area di circolazione. Questo caso è il duale del caso di acquisizione del territorio. Quindi, un comune può cedere un insieme di aree di circolazione (se l'insieme delle aree di circolazione costituisce un nuovo comune determina l'evento di nascita del comune stesso); una provincia può cedere uno o più comuni (se l'insieme dei comuni costituisce una nuova provincia determina l'evento di nascita della provincia stessa), e, infine, un'area di circolazione può cedere una sua porzione ad un'altra area di circolazione.

La Figura 5.6 illustra come le componenti del dato toponomastico puro interessate da *tutte le cinque tipologie di eventi di origine* sono: provincia, comune, area di circolazione. Invece, il numero civico (non rappresentato in Figura 5.6) è interessato solamente da eventi di nascita, cambio di denominazione e scomparsa.

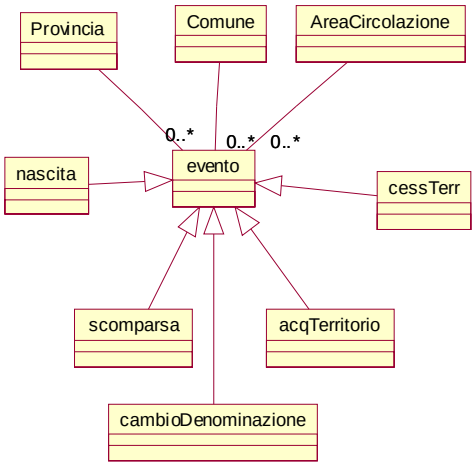


Figura 5.6: Tipologie di eventi di origine e componenti del dato toponomastico puro cui gli eventi possono riferirsi.

5.2.4 Progettazione dei Flussi

In questo Paragrafo si descrive la progettazione dei flussi di scambio dei dati toponomastici, che utilizza l'analisi e le classificazioni effettuate nei paragrafi precedenti. In primo luogo, nel Paragrafo 5.2.4.1 sono presentate alcune caratteristiche dei dati toponomastici che possono essere sfruttate per la progettazione dei flussi. Nel Paragrafo 5.2.4.2 viene descritto lo schema utilizzato per la progettazione dei flussi. Infine, nel Paragrafo 5.2.4.3, sono descritti alcuni esempi di applicazione dello schema di progettazione proposto.

E' importante notare come lo sforzo di generalizzazione descritto nei paragrafi 5.2.4.1 e 5.2.4.2 è estremamente utile soprattutto nell'ottica di modifiche future che possono essere fatte sulla struttura degli indirizzi gestiti dalle PA. La generalità dello schema di progetto dei flussi proposto ne consente, infatti la pressocchè immediata applicazione ad eventuali nuove componenti degli indirizzi e quindi il ridisegno dei flussi connessi può avvenire in maniera diretta.

5.2.4.1 La Struttura Gerarchica del Dato Toponomastico e la Proprietà di Ereditarietà degli Eventi

Il dato toponomastico puro presenta una peculiarità strutturale che consiste nell'organizzazione a livelli gerarchici. In particolare, le componenti del dato toponomastico possono essere organizzate sulla base di una gerarchia a livelli, tale per cui l'elemento a livello i+1 *contiene* l'elemento a livello i. Tale proprietà è stata precedentemente citata, ed è rappresentata nella Figura 4.2. Per comodità di lettura si riporta la gerarchia anche in Figura 5.7.

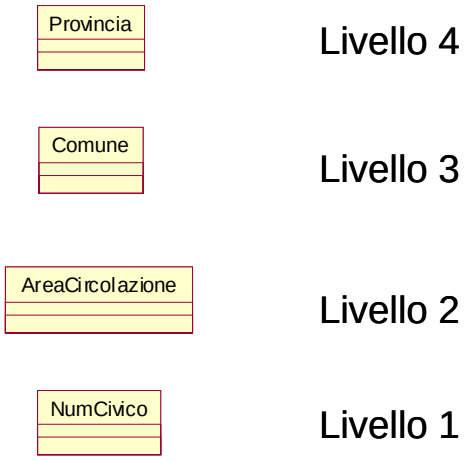


Figura 5.7: La struttura a livelli del dato toponomastico

Tra i livelli della gerarchia esiste un'interessante *proprietà di ereditarietà* legata agli eventi. Informalmente, le informazioni ai vari livelli della gerarchia non sono tipicamente memorizzate su supporto informatico senza le informazioni di livello più alto, perché necessitano di queste ultime per essere identificate. Cosi' i comuni necessitano della provincia, le aree di circolazione dei comuni e delle province, e i numeri civici delle aree di circolazione, dei comuni e delle province. Questo comporta che un evento di origine a livello i+1 della gerarchia sia ereditato dai livelli sottostanti (qualora sussista il vincolo di identificazione descritto).

La proprietà di ereditarietà è estremamente utile nella progettazione dei flussi come verrà descritto nel Paragrafo successivo.

5.2.4.2 Pattern per la Riprogettazione dei Flussi per il Miglioramento della Qualità dei Dati

Sulla base delle classificazioni descritte nei paragrafi precedenti e della proprietà di ereditarietà degli eventi, è stato possibile disegnare un **Pattern Generale,** che fornisce la struttura reingegnerizzata di tutti i flussi di scambio dei dati toponomastici. Da un punto di vista metodologico, l'introduzione di un *pattern* generale ha un duplice vantaggio:

– in primo luogo, consente un ridisegno uniforme dei flussi tra le varie pubbliche amministrazioni. L'omogeneità dei flussi progettati ha un impatto positivo sulla fase di *change management*, che segue gli interventi di reingegnerizzazione dei processi, in quanto semplifica l'apprendimento dei nuovi flussi in cui ciascuna amministrazione si trova ad essere inserita;

– in secondo luogo, come accennato, facilita l'estensibilità futura dell'insieme dei flussi progettati.

Il *pattern* è mostrato in Figura 5.10. X rappresenta una delle compoenenti del dato toponomastico puro identificate in Figura 5.2.

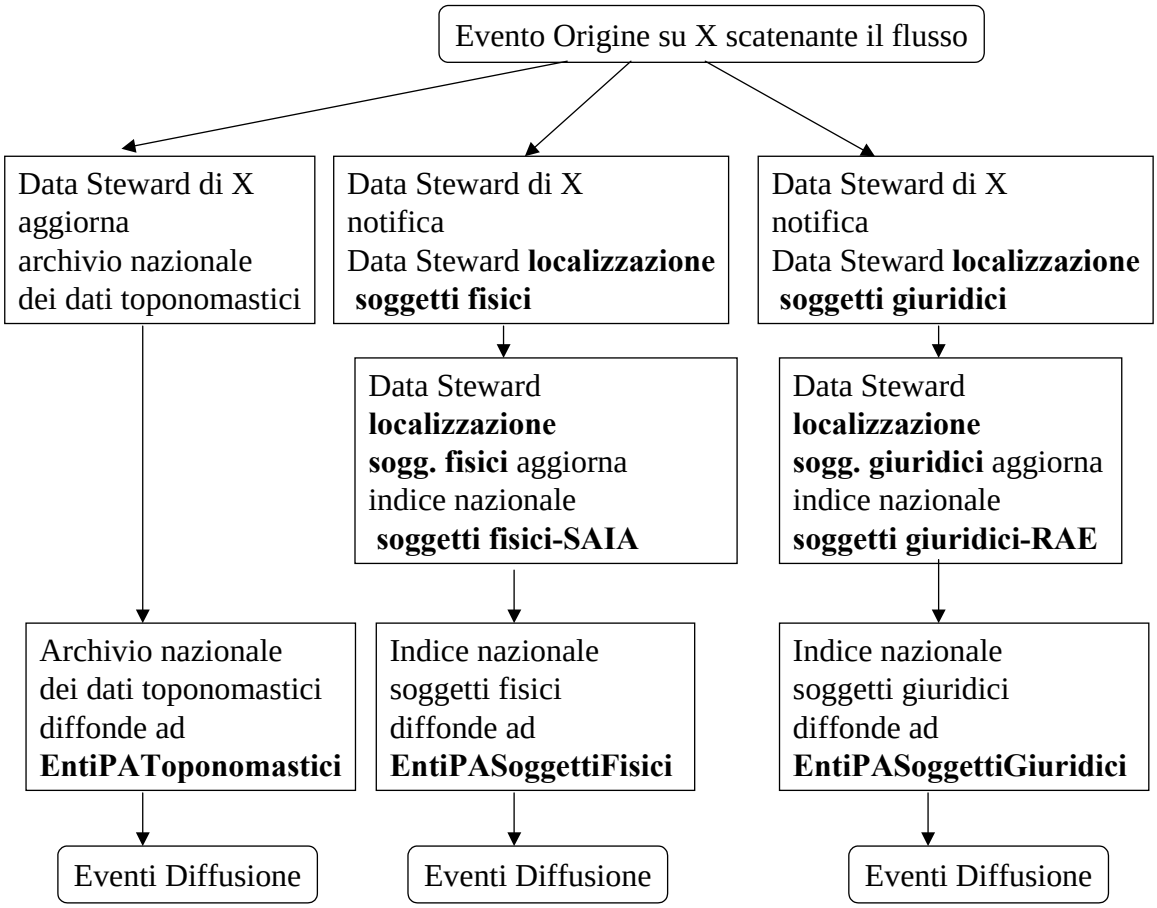


Figura 5.10: Pattern generale suggerito per la progettazione dei flussi

L'idea di base è che dovunque si determini un evento origine su X, il *Data Steward* di X ne deve essere notificato. E' a tal punto compito del *Data Steward* di X attivare una propagazione "guidata" degli eventi di diffusione. In particolare, non appena si verifica un evento di origine su X, il *Data Steward* di X provvede all'aggiornamento del nuovo valore di X presso l'archivio nazionale dei dati toponomastici; quindi, eventi di diffusione notificano agli enti PA toponomastici l'avvenuto cambiamento (blocchi a sn della figura).

Nella parte centrale della figura, il *Data Steward* di X notifica al *Data Steward* di localizzazione dei soggetti fisici l'avvenuta variazione di X (evento di diffusione) e quest'ultimo provvede alla notifica della variazione di X agli enti PA Soggetti Fisici mediante eventi di diffusione.

Infine, a destra della figura è descritta una sequenza analoga a quella appena descritta, nel caso dei soggetti giuridici: il *Data Steward* di X notifica al *Data Steward* di localizzazione dei soggetti giuridici l'avvenuta variazione di X (evento di diffusione)

e quest'ultimo provvede alla notifica della variazione di X agli enti PA Soggetti Giuridici mediante eventi di diffusione.

Inoltre sfruttando la proprietà di ereditarietà, il *pattern* generale appena illustrato può essere completato applicando i seguenti passi:

1. sia O l'evento origine su X. Supponiamo che X sia il dato j-esimo del livello i+1, ed indichiamolo con $X^{(J)}_{(i+1)}$
2. se vale la *proprietà di ereditarietà* per ogni dato ai livelli k<i+1, l'evento O viene ereditato da ogni $X^{(J)}_{(k)}$
3. è quindi possibile applicare il *pattern* generale per costruire i flussi generati da ogni $X^{(J)}_{(k),}$ con k<i+1.

I passi descritti considerano l'ipotesi generale di un numero di componenti maggiore di uno al livello i+1 del dato toponomastico. Invece, in Figura 5.7 si considerano singole componenti per ogni livello. Ovviamente, i passi descritti si estendono in maniera immediata al caso di singola componente per livello. Ad esempio, per un evento di origine sui comuni, gli eventi di diffusione dovranno interessare oltre agli archivi che memorizzano i comuni, anche quelli che memorizzano aree di circolazione e numeri civici.

Gli interventi a livello di singoli archivi, che devono essere intrapresi una volta ricevuti gli eventi di diffusione, dipendono da come i dati toponomastici sono logicamente e fisicamente organizzati. Se, ad esempio, si ha un cambio di denominazione di un comune, e un'amministrazione memorizza una tabella dei comuni ed un tabella con indirizzi completi (di cui il comune è uno dei campi), gli eventi di diffusione iniziali interesseranno l'aggiornamento della sola tabella dei comuni, mentre gli eventi di diffusione successivi interesseranno l'aggiornamento della tabella degli indirizzi completi.

Il processo generale per la costruzione di tutti i flussi originati dall'evento O è schematizzato in Figura 5.11.

5.2.4.3 Benefici in Termini di Miglioramento della Qualità dei Dati

Il ridisegno dei flussi proposto è guidato dalla necessità di avere flussi di interscambio dei dati toponomastici che ne garantiscano la qualità. In particolare, i benefici conseguiti in termini di miglioramento della qualità dei dati sono di seguito descritti:

- miglioramento dell'aggiornamento dei dati toponomastici. I dati toponomastici sono caratterizzati da una variabilità temporale, che rende difficile assicurare l'aggiornamento di medesime copie dello stesso dato, memorizzate da differenti pubbliche amministrazioni. Gli eventi di diffusione previsti dal *pattern* generale attivano un flusso di notifiche definito in maniera precisa, che consente di risolvere la problematica dell'aggiornamento;
- miglioramento dell'accuratezza sintattica e della completezza. L'adozione effettiva di un formato di interscambio automatico, è subordinata ad un'adeguata riprogettazione dei flussi che possa supportare l'interscambio automatico. La riprogettazione suggerita, qualora fosse implementata, potrebbe adeguatamente supportare un interscambio automatico dei dati toponomastici con i benefici che

ne possono conseguire, sia in termini di accuratezza sintattica che di completezza. In termini di accuratezza sintattica, un notevole miglioramento deriverebbe dall' acquisizione automatica dei dati trasmessi. In termini di completezza, l'implementazione di un interscambio automatico consentirebbe di effettuare controlli non solo in fase di acquisizione automatica o da utente umano, nell'ambito dei flussi intra-amministrazione, ma anche nel caso di flussi inter-amministrazioni (si veda anche il Paragrafo 3 del Capitolo 4).

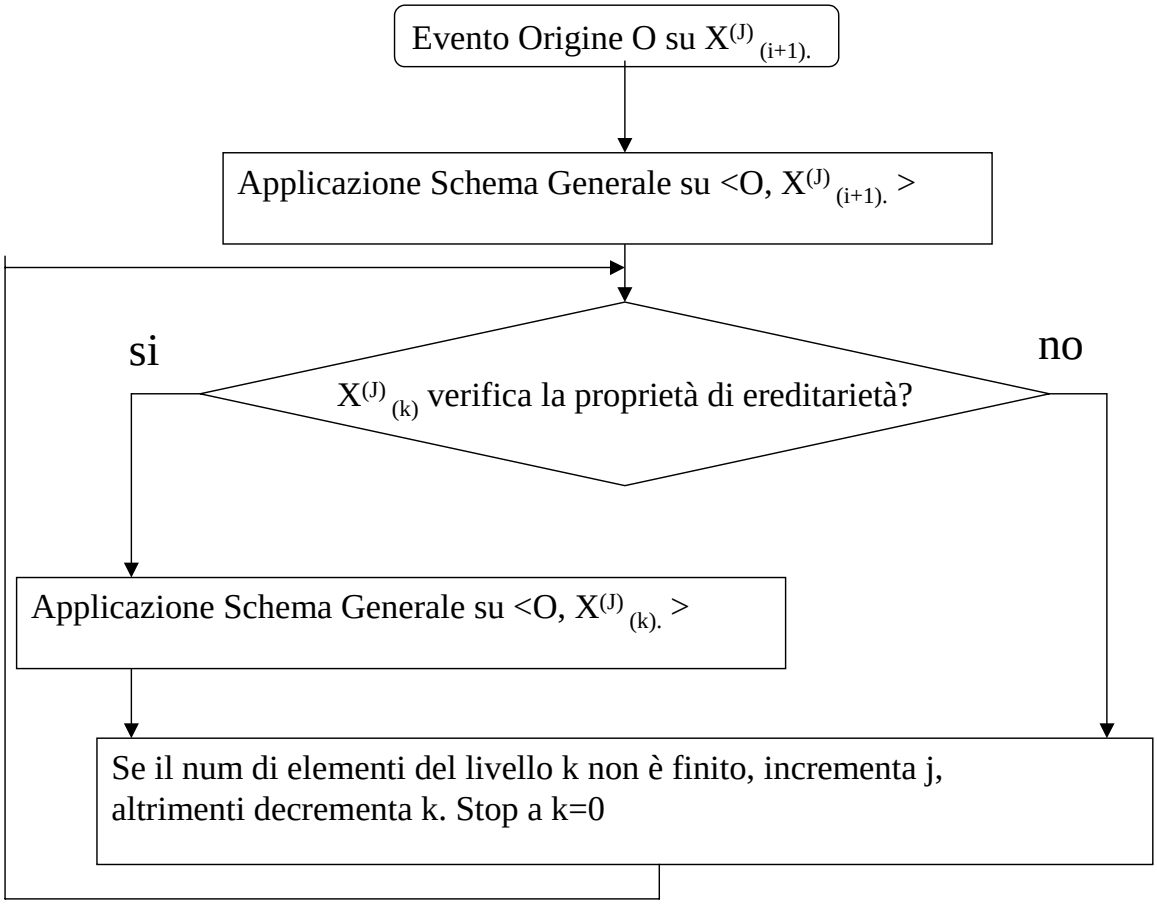


Figura 5.11: Processo generale per la costruzione dei flussi

5.2.4.4 Esempio 1. Variazione di Denominazione di un Comune

Evento di origine: Variazione di denominazione

Componente Dato Toponomastico Puro: Comune

Le Figure 5.12, 5.13 e 5.14 mostrano l' applicazione del *pattern* generale di progetto con O=Variazione di denominazione ed X=Comune.

L' applicazione della proprietà di ereditarietà implica che l'implementazione del metodo `CambiaComune,` in ciascuno dei diagrammi mostrati nelle Figure 5.12, 5.13 e 5.14, comporti l'aggiornamento del campo Comune laddove sono memorizzate le Aree di circolazione e i Numeri Civici. Come esempio, è riportato in Figura 5.15 l'implementazione del metodo CambiaComune realizzato dall'Archivio Nazionale Toponomastico.

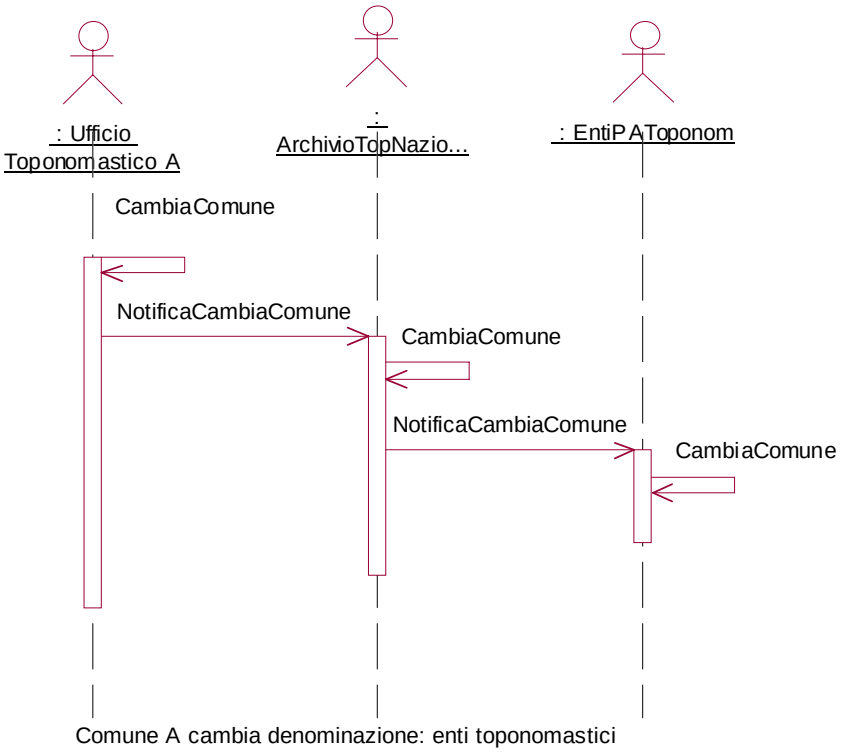


Figura 5.12: Eventi di diffusione ad enti PA toponomastici.

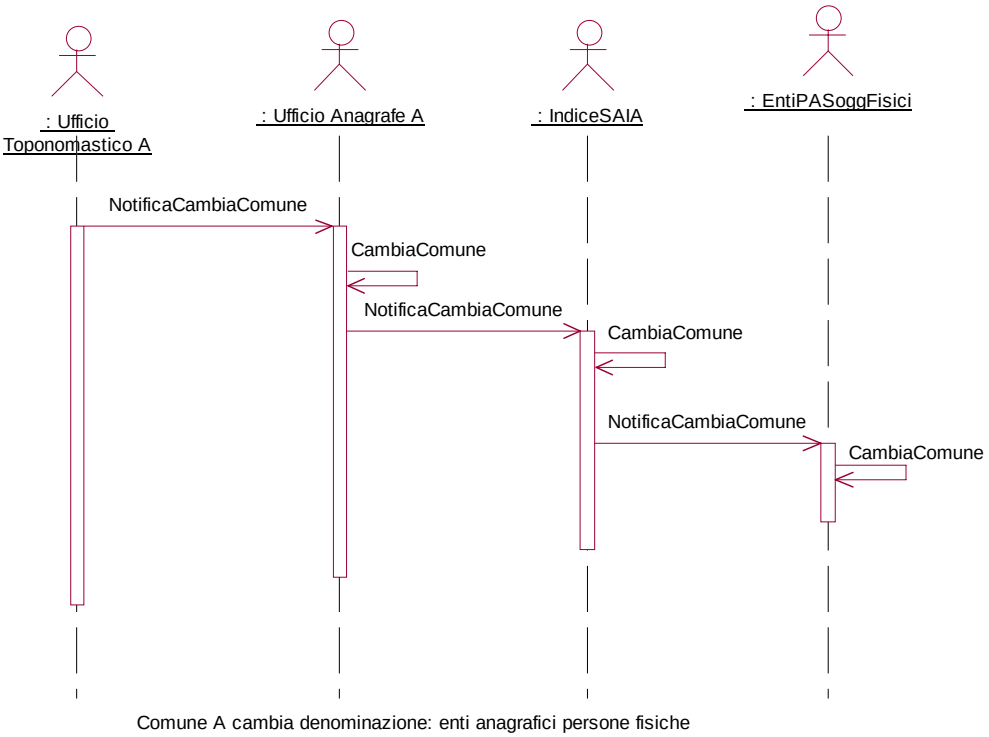


Figura 5.13: Eventi di diffusione ad Enti PA Soggetti Fisici.

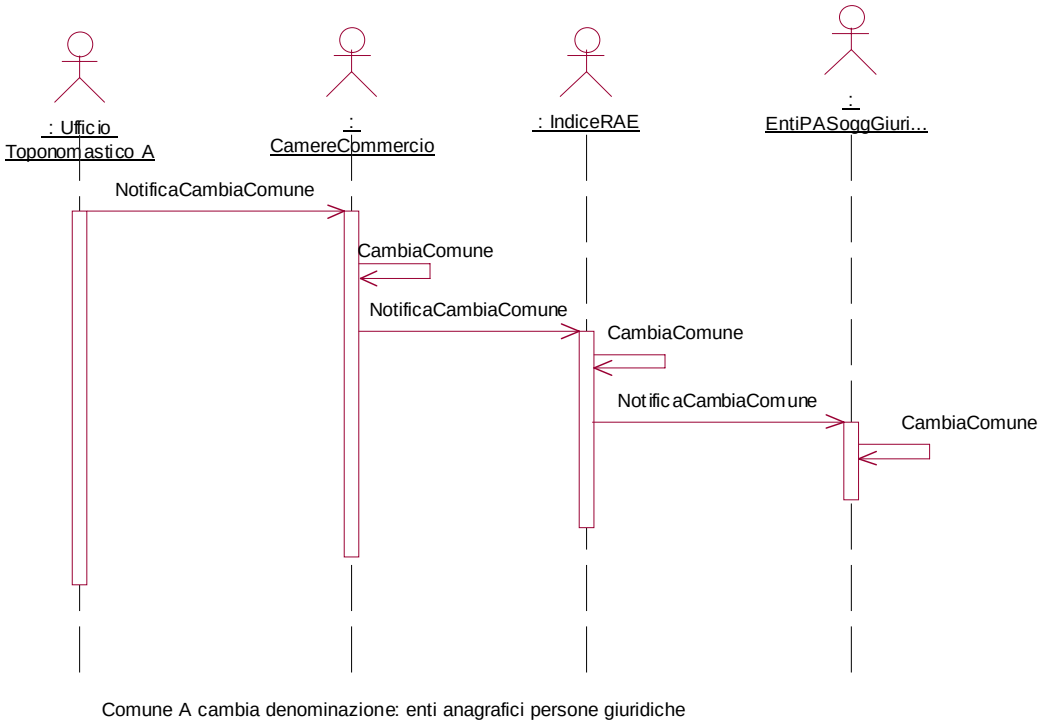


Figura 5.14: Eventi di diffusione ad Enti PA Soggetti Giuridici.

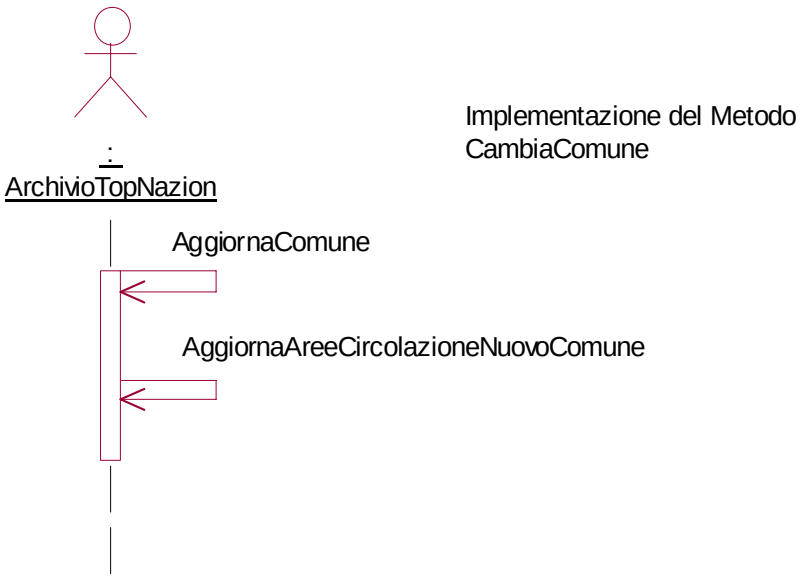


Figura 5.15: Aggiornamento delle Aree di Circolazione con il nuovo comune come evento ereditato dal cambio di denominazione del comune.

5.2.4.5 Esempio 2. Acquisizione Aree di Circolazione: Comune A acquisisce da Comune B

Evento di origine: Acquisizione senza scomparsa

Componente Dato Toponomastico Puro: Aree di Circolazione

Le Figure 5.16, 5.17 e 5.18 mostrano l' applicazione del pattern generale di progetto con O=Acquisizione senza scomparsa ed X=Aree di Circolazione.

Si noti anche come l' applicazione della proprietà di ereditarietà implica che l'implementazione del metodo `AggiornaAree di Circolazione,` in ciascuno dei diagrammi mostrati nelle Figure 5.16, 5.17 e 5.18, comporti l'aggiornamento del campo Comune laddove sono memorizzate i Numeri Civici corrispondenti alle aree di circolazione di interesse.

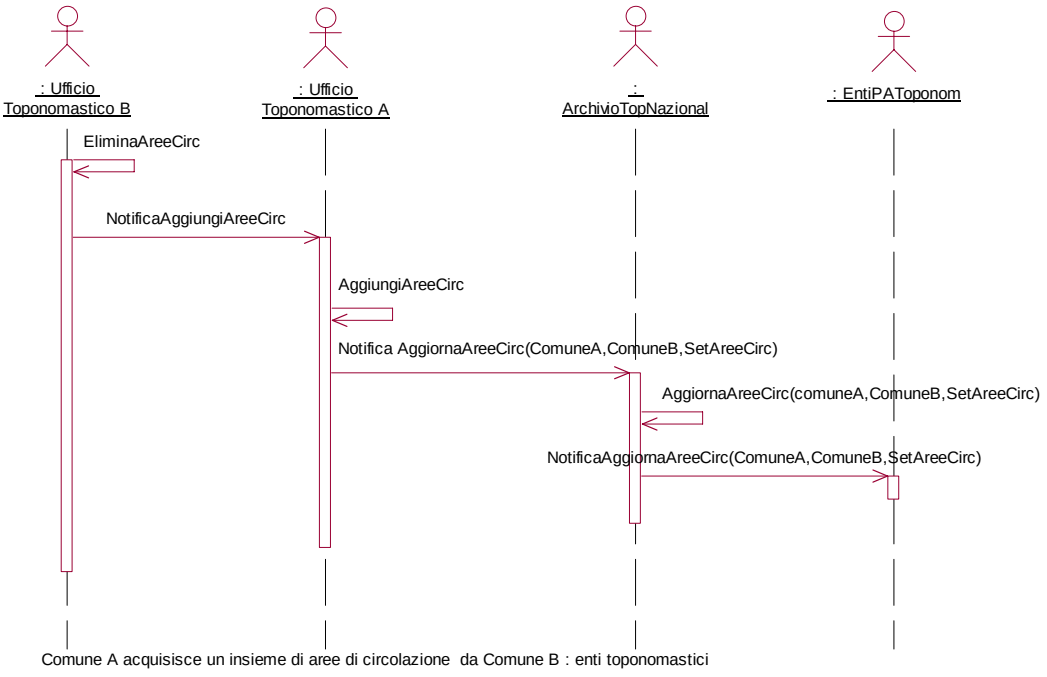


Figura 5.16: Eventi di diffusione ad enti PA toponomastici.

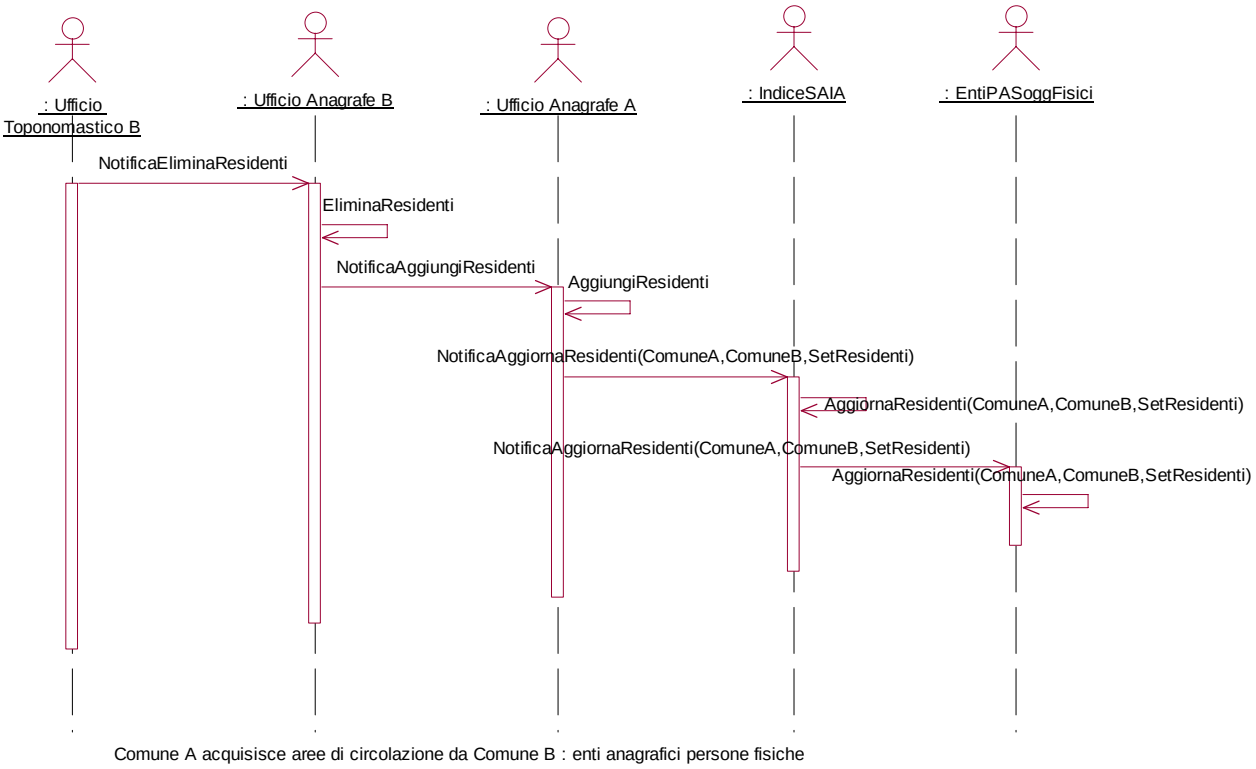


Figura 5.17: Eventi di diffusione ad Enti PA Soggetti Fisici.

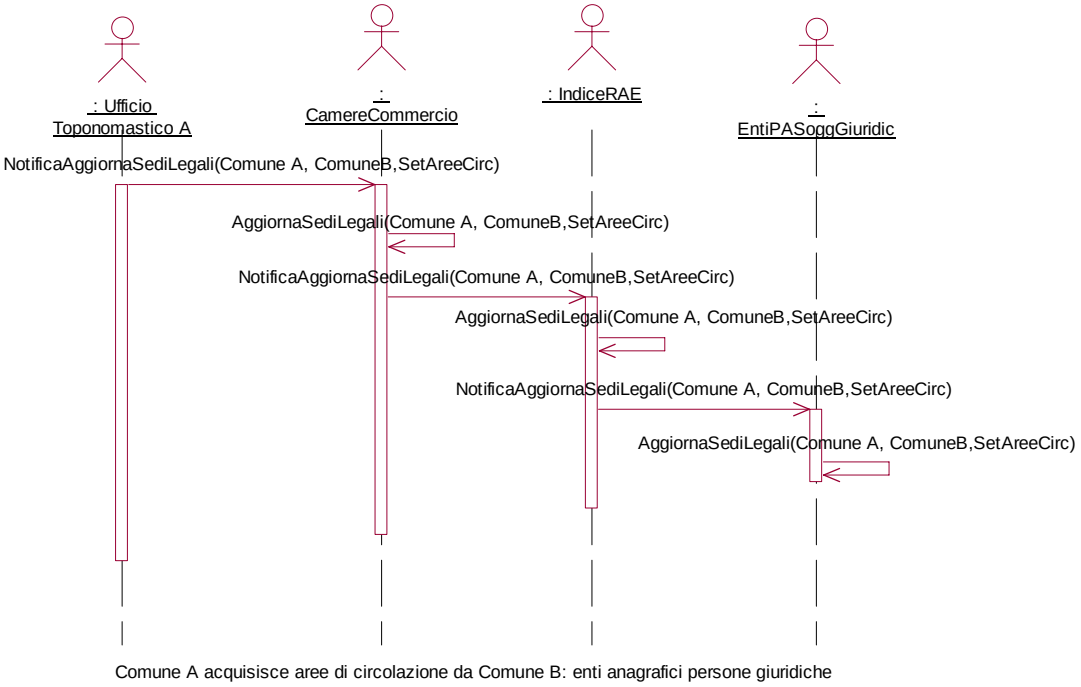


Figura 5.18: Eventi di diffusione ad Enti PA Soggetti Giuridici.

6 Conclusioni e Sviluppi Futuri

6.1 Sintesi dei Risultati

Il presente volume illustra l'insieme dei *principi guida* che le pubbliche amministrazioni italiane possono seguire al fine di *migliorare la qualità delle proprie basi di dati toponomastiche.* Tali principi guida derivano dagli studi e dalle esperienze condotte nell'ambito dell'accordo di collaborazione, per la definizione di criteri guida, per la gestione della qualità dei dati nella pubblica amministrazione, stipulato tra l'Autorità per l'Informatica nella Pubblica Amministrazione (AIPA, ora CNIPA) e l'ISTAT.

Le componenti fondamentali dei principi guida proposti sono:

- la definizione di formati di acquisizione ed interscambio degli indirizzi, che consentono alle amministrazioni di effettuare controlli di qualità prima di memorizzare gli indirizzi, sia nella fase di acquisizione diretta dal cittadino, che nella fase di acquisizione e scambio con un'altra amministrazione;
- la proposta di intervento sui processi di scambio degli indirizzi, con l'obiettivo principale di garantire l'allineamento e l'aggiornamento degli indirizzi nelle basi di dati pubbliche;
- la definizione di una metodologia di misurazione, basata sulla definizione di indicatori e tecniche statistiche, per la valutazione della qualità delle basi di dati toponomastiche, a vari livelli di complessità e sulla base di differenti requisiti.
- l'ideazione di un processo che consente di integrare le tecniche di misura e il miglioramento basato sui processi e la definizione di formati standard, nell'ottica di una più elevata qualità delle basi di dati toponomastiche, che le amministrazioni pubbliche possono seguire.

Il lavoro condotto è particolarmente rilevante, in quanto, in base alla nostra conoscenza, è una delle prime iniziative ad investigare in maniera profonda una particolare tipologia di dati, memorizzati dalle amministrazioni pubbliche, ovvero i dati toponomastici. Nello specifico, si è condotta non solo un'analisi teorico-concettuale delle problematiche relative ai dati toponomastici, ma sono stati anche investigati gli aspetti normativi connessi all'utilizzo effettivo di tali dati, tale da poter suggerire, in maniera adeguata, i principi guida per garantire la qualità dei dati toponomastici nelle amministrazioni pubbliche. Inoltre, le sperimentazioni condotte su alcuni archivi nazionali, hanno consentito di capire la natura e le caratteristiche dei problemi di qualità dei dati toponomastici ed hanno guidato l'ideazione delle strategie di miglioramento, suggerite nell'ambito dei principi guida.

Il processo proposto alle amministrazioni, per migliorare la qualità dei propri dati toponomastici, rispetta alcune caratteristiche fondamentali per interventi che riguardano amministrazioni pubbliche. Infatti, le azioni di miglioramento sono state concepite in maniera tale da limitare il più possibile cambiamenti interni alle

amministrazioni stesse e nell'ottica di salvaguardare il patrimonio informativo pre-esistente. Ad esempio, la selezione delle componenti, che definiscono gli indirizzi, è avvenuta a partire dall'esame della struttura di sei archivi nazionali di dati toponomastici, appartenenti sia ad enti pubblici che ad enti privati (vedi Capitolo 2, Paragrafo 2.1.2); la selezione delle componenti stesse è stata effettuata: sulla base delle caratteristiche strutturali degli indirizzi presi in esame, nel rispetto del fondamento normativo e in funzione del principio di minima ridondanza.

Alcuni dei principi guida proposti sono di tipo specifico per i dati toponomastici, ad esempio le azioni di intervento sui processi sono intrinsecamente legate alla natura del dato toponomastico. Tuttavia, l'approccio metodologico al problema di migliorare la qualità dei dati toponomastici può essere applicato anche ad altre tipologie di dati. In particolare, il processo di miglioramento proposto, che comprende le fasi di misurazione, miglioramento interno e miglioramento basato sulla cooperazione, ha caratteristiche di generalità tali da poter essere adattato anche al miglioramento di altre tipologie di dati. Inoltre, la fase di miglioramento basata sulla cooperazione, che nel caso degli indirizzi è stata fondamentalmente incentrata sul problema dell'aggiornamento e dell'allineamento delle basi di dati toponomastiche, può includere tecniche di *record linkage*, la cui applicazione può avere valenza del tutto generale.

6.2 Sviluppi Futuri

I risultati conseguiti sono la base di possibili attività future finalizzate alla realizzazione di ulteriori prodotti, fondamentali all'effettiva realizzazione dei principi guida introdotti.

Si suggerisce in primo luogo di avviare un processo, che conduca alla costruzione di un archivio nazionale degli indirizzi. Nel presente volume, l'esigenza di un archivio nazionale degli indirizzi è emersa più volte. Diversi contributi possono concorrere alla realizzazione effettiva di tale archivio, in particolare:

- la definizione delle componenti dell'indirizzo e la loro rappresentazione nei formati standard;
- la fornitura di vocabolari standard di riferimento per alcune componenti degli indirizzi, riportati in Appendice;
- la definizione di possibili flussi di alimentazione dell'archivio, nel contesto dei processi di aggiornamento dell'informazione toponomastica.

Per arrivare alla realizzazione effettiva dell'archivio si potrebbe seguire un processo in cui: *(i)* si costruisce una prima versione dell'archivio a partire da archivi nazionali esistenti, *(ii)* si valida tale versione coinvolgendo in modo strutturato i Comuni che costituiscono *data steward* per alcune componenti dell'indirizzo.

Per la costruzione della prima versione dell'archivio, si potrebbe partire dallo Stradario Nazionale realizzato dall'Istat in occasione degli ultimi Censimenti Generali della Popolazione, in base alle indicazioni fornite dai comuni. Mediante l'adozione di algoritmi di *Record Linkage* tale archivio può essere incrociato con altri archivi nazionali di dati toponomastici come, ad esempio, con l'archivio dei dati

toponomastici detenuto da Poste Italiane. Questa operazione permetterebbe di utilizzare versioni diverse del medesimo indirizzo e di ottenere, pertanto, una qualità ottimale dell'archivio nazionale risultante. Gli algoritmi di *record linkage* possono essere altresì utilizzati nella fase di validazione dell'archivio con i Comuni.

6.2.1 Ringraziamenti

Desideriamo ringraziare in primo luogo Carlo Batini e Francesco Zannella, che hanno reso possibile la realizzazione dell'accordo, nell'ambito del quale il presente lavoro si colloca.

Un ringraziamento particolare ai coordinatori scientifici Alessandro Alessandroni e Piero Demetrio Falorsi.

Un ringraziamento particolare a Monica Scannapieco per il contributo fornito in merito agli aspetti informatici ed ingegneristici.

Desideriamo anche ringraziare Elettra Cappadozzi che ha fornito preziose indicazioni in merito ai formati standard proposti, Gabriele Ciasullo per il supporto fornito all'analisi della normativa vigente, ed Enrica Massella per le indicazioni ed i commenti sul ridisegno dei flussi di interscambio.

Ringraziamo Poste Italiane per la gentile collaborazione e per il supporto fornito alla definizione dei formati.

Un ringraziamento sentito a tutto il gruppo di lavoro ISTAT, che ha consentito il conseguimento dei risultati descritti nel presente volume: Antonia Boggia, Fabio Crescenzi, Marcello D'Orazio, Orietta Gargano, Alessandro Pallara, Antonio Pavone, Lamberto Pizzicanella, Marina Signore, Giorgia Simeoni

Bibliografia

[AIPA 2002] I dati pubblici; linee guida per l'accesso, la comunicazione e la diffusione. Quaderni AIPA, Aprile 2002

[Bilenko2003] Bilenko, M. and Mooney, R. J., Adaptive Duplicate Detection Using Learnable String Similarity Metrics, Proceedings of ACM Conference on Knowledge Discovery and Data Mining, Washington, DC, August 2003, 39-48.

[BrackstoneGordon1999] Brackstone and Gordon: Managing Data Quality in a Statistical Agency, Statistics Canada, Survey Methodoloy, Catalogue N° 12-001-XBP, Vol. 25 N° 2, 1999. Disponibile a: http://dsbb.imf.org/scpap.pdf.

[Breiman1984] Breiman L., Friedman J.H., Olshen R.A., Stone C.: Classification and regression trees, Wadsworth, Belmont, California, 1984.

[CIS] J.Mylopoulos and M.P.Papazoglou (eds.), Cooperative Information Systems (Special Issue), IEEE Expert Intelligent Systems & Their Applications, vol. 12, no. 5, 1997.

[Cicchitelli1992] Cicchitelli G., Herzel A., Montanari G. E., Il campionamento statistico, Il Mulino, 1992.

[Cohen2003] Cohen, W. W., Ravikumar, P., and Fienberg, S. E. (2003a), A Comparison of String Metrics for Matching Names and Addresses," International Joint Conference on Artificial Intelligence, Proceedings of the Workshop on Information Integration on the Web, Acapulco, Mexico, August 2003.

[Eurogeographics] Eurogeographics: http://www.eurogeographics.org

[Eurostat1999] Eurostat: Quality Work and Quality Assurance within Statistics, 88th DGINS Conference,1999.

[Eurostat2000] Eurostat: Standard quality report, 2000. Disponibile a : http://www.forum.europa.eu.int/, 2000.

[Eurostat2001A] Eurostat: How to make a quality report, version 2,2001.

[Eurostat2001B] Eurostat: LEG Quality, CPS 2001/42/7, 2001.

[IntesaGIS] IntesaGIS: http://www.intesagis.it/.

[INSPIRE] INSPIRE : http://www.ec-gis.org/inspire/

[ISTAT1992] Anagrafe della popolazione, legge e regolamento anagrafico. Metodi e norme. Serie B – n. 29, ISTAT, 1992.

[Navarro2001] Navarro G, A guided tour to approximate string matching, ACM Computing Surveys (CSUR), Vol. 33 , No. 1, March 2001.

[RAE] Bertoletti M., Missier P., Scannapieco M., Aimetti P., Batini C.: Improving Government-to-Business Relationships through Data Reconciliation and Process Re-engineering. To appear on Advances in Management Information Systems-

Information Quality Monograph (AMIS-IQ) Monograph (Richard Wang, ed.), Sharpe, M.E., 2004.

[Redman1996] Redman T.C: .Data Quality for the Information Age. Artech House, 1996.

[SAIA] SAIA: http://www.servizidemografici.interno.it/sitoCNSD/

[Shankaranayan2000] Shankaranayan G., Wang R. Y. and Ziad M.: Modeling the Manufacture of an Information Product with IP-MAP. In Proceedings of the 6th International Conference on Information Quality, Boston, MA, 2000.

[SAS2004] SAS, http: //www.sas.com/offices/europe/italy/index.html/

[SPSS2004] SPSS, http://www.spss.it.

[SistemiInformativi2001] G. Lazzi, Reingegnerizzazione dei processi in Batini et al.: Sistemi informativi, Franco Angeli, volume 1 capitolo 3, 2001.

[StatisticsCanada1998] Statistics Canada: Quality guidelines, Third Edition,1998.

[UML2004] OMG document: Unified Modeling Language vs.1.5. Disponibile a : http://www.omg.org/technology/documents/formal/uml.htm.

[Wang1995] Wang R.Y., Storey V.C., Firth C.P: A Framework for Analysis of Data Quality Research. IEEE Transaction on Knowledge and Data Engineering, Vol. 7, No. 4, 1995.

[WangStrong1996] Wang R.Y., Strong D.M., Beyond Accuracy: What Data Quality Means to Data Consumers. Journal of Management Information Systems, vol. 12, no. 4, 1996.

[Winkler2004] William E. Winkler Methods for evaluating and creating data quality, Information Systems, Volume 29, Numero 7, pagg. 529-636, 2004.

[Winkler2004] Winkler, W. E. (1994), Advanced Methods for Record Linkage, Proceedings of the Section on Survey Research Methods, American Statistical Association, 467-472 (longer version report 94/05 available at http://www.census.gov/srd/www/byyear.html).

[Wolter1985] Wolter K.M., Introduction to Variance Estimation, Springer-Verlag, New York, 1985.

[XML2004] T.Bray, J.Paoli, C.M.Sperberg, E.Maler, F.Yergeau: eXtensible Markup Language (XML). Version 1.0 (Third Edition), http://www.w3.org/TR/2004/REC-xml-20040204/, February 2004.

[XMLSchema1] H.S. Thompson and D. Beech and M. Maloney and N.Mendelsohn: XML schema part 1: Structures, W3C Recommendation. Disponibile alla URL: http://www.w3.org/TR/xmlschema-1/, Maggio 2001.

[XMLSchema2]P.V. Biron and A.Malhotra: XML schema part 2: Datatypes, W3C Recommendation. Disponibile alla URL: http://www.w3.org/TR/xmlschema-2 , Maggio 2001.

Appendice 1. Glossario

Area di circolazione
Elemento lineare o areale, di qualsiasi forma o misura, destinato alla viabilità (via, strada, corso, viale, vicolo, calle, salita, piazza, piazzale, largo, campiello e simili…).

Arco di strada
Tratto di strada compreso tra due intersezioni stradali, identificato da una coppia di numeri civici (civico iniziale e civico finale).

Base dati
Insieme di informazioni di diverso tipo, organizzate secondo criteri ben precisi (modello logico dei dati) in un qualsiasi sistema gestionale standard (modello fisico).

Codici territoriali
Codici identificativi degli enti amministrativi territoriali (regioni, province, comuni) e delle sezioni di censimento. Si fa riferimento ai codici ufficiali dell'Istat.

Dati toponomastici
Dati relativi ai toponimi delle varie entità territoriali e/o amministrative che possono essere definite sul territorio (nomi, codici, ecc).

Dug (Denominazione Urbanistica Generica)
Tipologia dell'area di circolazione. Esistono "n" tipologie possibili come ad esempio via, piazza, località, contrada, ecc. Nella pratica corrente sono note definizioni diverse, come "specie", "tipo", "particella toponomastica".

Frazione geografica
E' costituita da un'area di territorio comunale comprendente di norma un centro abitato, nonché nuclei abitati e case sparse gravitanti sul centro.

Georeferenziazione
Procedura che consiste nel posizionare, mediante l'ausilio di punti a coordinate note detti punti di controllo, dati vettoriali o raster nella rispettiva zona del territorio reale secondo un determinato sistema di riferimento.

Geocodifica
Assegnazione di codici territoriali ad un oggetto come ad esempio un indirizzo.

Indirizzo
Sequenza di elementi toponomastici organizzati gerarchicamente che identificano in modo univoco un luogo fisico sul territorio.

Itinerario di sezione
E' l'elenco delle aree di circolazione con gli intervalli dei numeri civici relativi che ricadono all'interno delle sezioni di censimento.

Località Istat
Area sub-comunale di territorio, conosciuta di norma con un nome proprio, sulla quale sono situate più case raggruppate o sparse; si distinguono due tipi di località: località abitate e località produttive.

Località abitata
Area più o meno vasta di territorio, conosciuta con un nome proprio sulla quale sono situate una o più case raggruppate o sparse. Si distinguono tre tipi di località abitate: centro abitato, nucleo abitato, case sparse.

Centro abitato
Aggregato di case contigue o vicine con interposte strade, piazze e simili, o comunque brevi soluzioni di continuità, caratterizzato dall'esistenza di servizi od esercizi pubblici (scuola, ufficio pubblico, farmacia, negozio o simili) costituenti la condizione di una forma autonoma di vita sociale, e generalmente determinanti un luogo di raccolta ove sogliono concorrere anche gli abitanti dei luoghi vicini per ragioni di culto, istruzione, affari e simili, approvvigionamento e simili, in modo da manifestare l'esistenza di una forma di vita sociale coordinata dal centro stesso.

Nucleo abitato
Località abitata, priva del luogo di raccolta che caratterizza il centro abitato, costituita da un gruppo di case contigue e vicine, con almeno cinque famiglie, con interposte strade, sentieri, piazze, aie, piccoli orti, piccoli incolti e simili, purché l'intervallo tra casa e casa non superi una trentina di metri e sia in ogni modo inferiore a quello intercorrente tra il nucleo stesso e la più vicina delle case manifestamente sparse.

Case sparse
Case disseminate per la campagna o situate lungo strade a distanza tale tra loro da non poter costituire nemmeno un nucleo abitato.

Località produttiva
Area in ambito extraurbano non compresa in centri e nuclei abitati nella quale siano presenti unità locali in numero superiore a 10, o il cui numero totale di addetti sia superiore a 200, contigue o vicine con interposte strade, piazze e simili, o comunque brevi soluzioni di continuità non superiori a 200 metri; la superficie minima deve corrispondere a 5 ettari.

Località postale
Zone sul territorio nazionale note con un toponimo proprio a cui corrispondono i circa 14.000 uffici postali.

Metadati
Insieme delle informazioni che consentono di conoscere il significato dei dati, il proprietario, la fonte, l'insieme dei valori ammessi, la validità temporale, eccetera.

Normalizzazione
Processo attraverso il quale vengono restituite in *output* le denominazioni di una località/comune e di un'area di circolazione contenute.

Numeri civici
Numeri di un'area di circolazione che contraddistinguono gli accessi alle singole unità ecografiche. Possono seguire l'ordine secondo la successione dei numeri naturali o secondo il sistema metrico.

Sezione di censimento
Unità territoriale minima del Comune utilizzata per la rilevazione censuaria a partire dalla quale sono ricostruibili per somma le varie unità geografiche ed amministrative di livello superiore. Ciascuna sezione di censimento deve essere completamente contenuta all'interno di una sola

località. Il territorio comunale deve essere esaustivamente suddiviso in sezioni di censimento; la somma di tutte le sezioni di censimento ricostruisce l'intero territorio nazionale.

Specie/tipo (riferito all'area di circolazione)
Vedi dug.

Stradario
Insieme di tutte le strade di un comune, una provincia o dell'intero territorio nazionale. E' previsto un codice identificativo di ciascuna strada.

Appendice 2. La normativa

Si illustrano le norme che disciplinano la localizzazione territoriale delle persone fisiche e delle unità economiche.
Le norme si riferiscono ad una serie di variabili regolamentate, per le quali il valore di riferimento è definito dalle norme stesse, come anche i soggetti responsabili della loro acquisizione.
In particolare le variabili prese in esame sono le seguenti:

- residenza anagrafica;
- domicilio fiscale;
- sede delle persone giuridiche;
- sede delle imprese.

Prima di entrare nel merito delle specifiche normative viene fornito un glossario di riferimento, al fine di rendere più chiara la lettura degli aspetti normativi che riguardano le variabili sopra elencate, delle quali nelle schede che seguono si rilevano inoltre, i soggetti responsabili dell'acquisizione e della modifica del dato che la variabile esprime, nonché i soggetti responsabili della conservazione del dato stesso nel relativo archivio.

Per quello che riguarda la localizzazione dei soggetti fisici e giuridici, è stata presa in esame la variabile toponomastica: "strada", intesa nella sua accezione più ampia (via, viale, vicolo, piazza, piazzale, ecc.), ed in particolare la "strada comunale".
Vengono inoltre forniti alcuni riferimenti normativi concernenti il Comune, quale ente locale dotato di propria autonomia, e le sue modificazioni territoriali.
Sono stati inoltre considerati gli aspetti normativi che riguardano i flussi informativi degli archivi contenenti i dati che si riferiscono alle suddette variabili regolamentate.

GLOSSARIO

Domicilio fiscale: il concetto di residenza fiscale può essere ricollegato, ove non sia possibile l'utilizzazione di altri criteri, al centro degli interessi vitali, ossia al luogo con il quale il soggetto ha il più stretto collegamento sotto il profilo degli interessi personali e patrimoniali.

Persona fisica: sono i singoli considerati nella loro individualità: tutte le persone fisiche hanno la personalità giuridica (artt. 22 Cost. e 1 C.C.) che si acquista al momento della nascita.

Persona giuridica: è un'entità astratta, cui l'ordinamento giuridico attribuisce la personalità giuridica se presenta determinati requisiti (es.: le società commerciali, gli enti pubblici)
Esistono, determinati enti che non hanno personalità giuridica (enti di fatto), ma che sono considerati dall'ordinamento *centri di imputazione di situazioni soggettive* (es.: partiti, sindacati).
Gli elementi essenziali che definiscono la persona giuridica sono:
- una organizzazione di persone e di mezzi;
- uno scopo per il cui perseguimento tale organizzazione si costituisce e "vive";
- il riconoscimento da parte dell'ordinamento, concesso con provvedimento dell'autorità su richiesta degli interessati o per legge.

Le persone giuridiche si possono classificare in:
- *associazioni* (o corporazioni) sono gruppi di persone che si associano per perseguire scopi comuni;
- *fondazioni* consistono in un insieme di beni materiali ai quali il proprietario (fondatore) spogliandosene, ha assegnato una particolare destinazione, in vista di uno scopo duraturo (anche oltre la durata della propria vita) e determinato;
- *pubbliche e private* (p.96 dir. amm.tivo)

Società: con il contratto di società, due o più persone conferiscono beni o servizi per l'esercizio in comune di un'attività economica allo scopo di dividerne gli utili (art. 2247 C.C.).

Impresa: è l'esercizio professionale di un'attività economica organizzata al fine della produzione o dello scambio di beni o di servizi *(tale definizione è tratta da quella di imprenditore così come definito dall'art. 2082 del C.C.).*

Azienda: è il complesso dei beni organizzati dall'imprenditore per l'esercizio dell'impresa (art. 2555 C.C.)[1]

Ditta: è il segno distintivo che contraddistingue la persona dell'imprenditore nell'esercizio dell'attività di impresa (nome commerciale). La ditta è mezzo di individuazione necessario in quanto, in mancanza di diversa scelta, essa coincide con il nome civile dell'imprenditore.

[1] Anche se nel linguaggio comune i termini "impresa" e "azienda" vengono spesso utilizzati indistintamente, giuridicamente assumono significati nettamente diversi (impresa come attività, azienda come complesso di beni). Tra azienda e impresa c'è, in realtà, un rapporto da mezzo a fine. In particolare il concetto di azienda attiene agli strumenti, o ai fattori che l'imprenditore utilizza nel processo produttivo.

1. RESIDENZA ANAGRAFICA

Definizione

Il Codice Civile individua le seguenti relazioni territoriale della persona fisica: dimora, residenza e domicilio.

L'art. 43 definisce la ***Residenza*** come *luogo in cui la persona ha la dimora abituale.*
Lo stesso articolo definisce il ***Domicilio*** come *luogo in cui la persona ha stabilito la sede principale dei suoi affari e interessi.*
Nei commenti a tale articolo viene precisato che *il domicilio è una situazione di diritto; non è pertanto necessario che il soggetto vi dimori,* mentre *la residenza è una situazione di fatto ed implica l'effettiva ed abituale presenza del soggetto in un dato luogo; può essere scelta e mutata liberamente.*

Archivio

L'art. 1 della Legge 24 dicembre 1954 n. 1228 (Ordinamento delle anagrafi della popolazione residente) stabilisce che *in ogni Comune deve essere tenuta* ***l'anagrafe*** *della popolazione residente.*
Nell'anagrafe della popolazione residente sono registrate le posizioni relative alle singole persone, alle famiglie ed alle convivenze, che hanno fissato nel Comune la residenza, nonché le posizioni relative alle persone senza fissa dimora che hanno stabilito nel comune il proprio domicilio.

Sulla base di quanto disposto dall'art. 3 della stessa legge 1228/54 *il Sindaco, quale ufficiale del Governo, è ufficiale dell'anagrafe*. Tale funzione può essere delegata al segretario comunale o ad altro impiegato del Comune.
Il successivo art. 4 dispone che *l'ufficiale d'anagrafe provvede alla regolare tenuta dell'anagrafe della popolazione residente ed è responsabile della esecuzione degli adempimenti prescritti per la formazione e la tenuta degli atti anagrafici.*

Iscrizione e Variazioni

La residenza anagrafica si costituisce con l'iscrizione nell'anagrafe del Comune di dimora abituale.
L'art. 2 della legge 1228/54 dispone che *è fatto obbligo ad ognuno di chiedere l'iscrizione nell'anagrafe del Comune di dimora abituale e di dichiarare alla stessa i fatti determinanti mutazioni di posizioni anagrafiche.*
In caso di trasferimento di residenza, da un Comune all'altro, è previsto l'obbligo di denuncia del cambiamento anche all'anagrafe del Comune di precedente residenza.

L'art. 15 del D.P.R. 30 maggio 1989, n. 223 (Nuovo regolamento anagrafico della popolazione residente), prevede che *qualora l'ufficiale d'anagrafe accerti che non siano state rese le dichiarazioni attinenti fatti che comportano l'istituzione o la mutazione di posizioni anagrafiche,* invita gli interessati a rendere dette dichiarazioni; in caso di mancata dichiarazione l'ufficiale stesso provvede direttamente ai necessari adempimenti.

Il citato D.P.R. 223/89 disciplina nel dettaglio le iscrizioni (art. 7) le mutazioni (art. 10) e le cancellazioni (art. 11) anagrafiche.

Anagrafe dei pensionati INPS

Con l'art. 34 della Legge 21 luglio 1965, n. 903 (Avviamento alla riforma e miglioramento dei trattamenti di pensione della previdenza sociale) *è istituita presso ciascun Comune l'anagrafe dei pensionati dell'Istituto Nazionale della Previdenza Sociale.*

2. DOMICILIO FISCALE

Definizione

Secondo l'art. 2 del DPR 22 dicembre 1986, n. 917 (Approvazione del testo unico delle imposte sui redditi*) ai fini delle imposte sui redditi si considerano residenti le persone che per la maggior parte del periodo di imposta sono iscritte nelle anagrafi della popolazione residente o hanno nel territorio dello Stato il domicilio o la residenza ai sensi del Codice Civile.*

Il domicilio fiscale non coincide con la residenza anagrafica nel caso di contribuenti all'estero, di variazione del Comune di residenza da meno di 60 giorni o di variazioni conseguenti a provvedimenti dell'Amministrazione finanziaria.

Archivio

Ai sensi dell'art. 1 del DPR 19 settembre 1973, n. 605 (Disposizioni relative all'anagrafe tributaria e al codice fiscale dei contribuenti) ***l'anagrafe tributaria*** *raccoglie e ordina su scala nazionale i dati e le notizie risultanti dalle dichiarazioni e dalle denunce presentate agli uffici dell'amministrazione finanziaria e dai relativi accertamenti, nonché i dati e le notizie che possono comunque assumere rilevanza ai fini tributari.*

Iscrizione e Variazioni

L'art. 2 del DPR 605/ 73 stabilisce che *sono iscritte all'anagrafe tributaria, secondo un sistema di codificazione stabilito con decreto del Ministero per le finanze, le persone fisiche, le persone giuridiche e le società, le associazioni ed altre organizzazioni di persone o di beni prive di personalità giuridica alle quali si riferiscono i dati e le notizie raccolti ai sensi dell'art. 1.*

Ai fini dell'attribuzione del numero di codice fiscale, sono richiesti (art. 4 dello stesso DPR) i seguenti dati:

- *per le persone fisiche: il cognome e il nome, il luogo e la data di nascita, il sesso e il domicilio fiscale;*
- *per i soggetti diversi da persone fisiche: la denominazione, la ragione sociale o la ditta, il domicilio fiscale.*

Nell'indicazione della sede e del domicilio fiscale devono essere specificati la via, il numero civico e il codice di avviamento postale.

3. SEDE DELLE PERSONE GIURIDICHE

Definizione

L'art. 46 del Codice Civile stabilisce che *quando la legge fa dipendere determinati effetti dalla residenza o dal domicilio, per le persone giuridiche si* fa riferimento *al luogo in cui è stabilita la loro sede.*

nota: *Occorre precisare che mentre la nozione di domicilio, inteso come centro di affari e di interessi ben si adatta alle persone giuridiche, non altrettanto può dirsi per la nozione di residenza.*

Nei commenti all'art. 46 del Codice Civile si rileva la seguente definizione di ***Sede effettiva*****:** *E' il luogo in cui le persone giuridiche ed, in generale, gli enti collettivi svolgono la propria attività. Non è sufficiente che vi sia uno stabilimento o una succursale, occorre infatti la presenza degli uffici degli amministratori e di coloro che hanno la rappresentanza dell'ente.*

Archivio

L'art. 33 del Codice Civile stabilisce che *in ogni provincia è istituito un pubblico* ***Registro delle persone giuridiche.***

In tale registro viene indicata (tra l'altro) la denominazione e la sede della persona giuridica, nonché il cognome e il nome degli amministratori.

Iscrizione e Variazioni

Il D.P.R. 10 febbraio 2000, n. 361 (Regolamento recante norme per la semplificazione dei procedimenti di riconoscimento di persone giuridiche private e di approvazione delle modifiche dell'atto costitutivo e dello statuto) dispone che (art. 1) *le associazioni, le fondazioni e le altre istituzioni di carattere privato acquistano la personalità giuridica mediante il riconoscimento determinato dall'iscrizione nel Registro delle persone giuridiche, istituito presso le Prefetture.*

L'iscrizione avviene mediante domanda presentata alla Prefettura nella cui provincia è stabilita la sede dell'Ente.

L'art 7 (Competenze delle Regioni e delle Province autonome) del DPR 361/2000 prevede che *il riconoscimento delle persone giuridiche private che operano nelle materie attribuite alla competenza delle regioni dall'art. 14 del DPR 24/7/77, n. 616, e le cui finalità statutarie si esauriscono nell'ambito di una sola regione, è determinato dall'iscrizione nel registro delle persone giuridiche istituito presso la stessa regione.*

L'art. 4 dello stesso D.P.R. 361/2000 prevede che *nel registro devono altresì essere iscritte le modificazioni dell'atto costitutivo e dello statuto, il trasferimento della sede e l'istituzione di* ***sedi secondarie.***

La prefettura provvede anche alla cancellazione dell'ente dal registro della persone giuridiche (art. 6 DPR 361/2000).

4. SEDE DELLE IMPRESE

Definizione

Nei commenti all'art. 2197 del Codice Civile si definiscono:

- ***Sede principale*** *- Il luogo nel quale l'imprenditore svolge la sua prevalente attività di direzione e di amministrazione dell'azienda, e non quello in cui si trovano i beni, né quello in cui risiede il professionista che ha l'incarico di trattare pratiche nell'interesse della società;*
- ***Sede secondaria*** *– Il luogo in cui operano rappresentanti dell'imprenditore, i quali pur essendo dipendenti economicamente e amministrativamente dalla sede principale, conservano un ampio ed autonomo potere di determinazione e di decisione*

Archivio

Con l'art. 8 della Legge 29 dicembre 1993, n. 580 (Riordinamento delle camere di commercio, industria, artigianato e agricoltura) viene *istituito presso la camera di commercio l'ufficio del registro delle imprese di cui all'articolo 2188 del Codice civile.*

Nei commenti all'art. 2188 del Codice Civile si definisce come **Registro delle imprese** lo *strumento di pubblicità legale non solo per le imprese collettive ma anche per le imprese commerciali individuali.*

Lo stesso art. 8 della L. 580/93 dispone che *l'ufficio provvede alla tenuta del registro delle imprese sotto la vigilanza di un giudice delegato dal presidente del tribunale del capoluogo di provincia* (giudice del registro).

Iscrizione e Variazioni

Le iscrizioni nel registro delle imprese sono eseguite su domanda sottoscritta dall'interessato (art. 2189 Codice Civile).

Qualora tale inscrizione non venga richiesta, il giudice del registro può ordinarla con decreto.

Secondo l'art. 2196 del Codice Civile al momento dell'iscrizione vanno indicati il cognome e il nome dell'imprenditore, la ditta, l'oggetto e la sede dell'impresa.

Lo stesso imprenditore deve chiedere *l'iscrizione delle modificazioni relative agli elementi suindicati e della cessazione dell'impresa.*

L'art. 2197 del Codice Civile dispone che in caso di istituzione di sedi secondarie l'iscrizione deve essere richiesta all'ufficio del registro delle imprese sia del luogo dov'è la sede principale che del luogo dove è istituita la sede secondaria.

L'iscrizione all'ufficio del registro delle imprese del luogo dov'è la sede principale deve essere richiesta anche nel caso di istituzione di sedi secondarie all'estero.

Sono soggette all'obbligo dell'iscrizione nel registro delle imprese le società e le società cooperative (art. 2200 Codice Civile) nonché gli enti pubblici che hanno per oggetto esclusivo o principale un'attività commerciale (art. 2201 Codice Civile).

Secondo l'art. 2 del D.P.R. 14 dicembre 1999, n. 558 (Regolamento recante norme per la semplificazione della disciplina in materia di registro delle imprese, nonché per la semplificazione dei procedimenti relativi alla denuncia di inizio attività e per la domanda di iscrizione all'albo delle imprese artigiane o al registro delle imprese per particolari categorie di attività soggette alla verifica di determinati requisiti tecnici) *sono iscritti in una sezione speciale del registro delle imprese gli imprenditori agricoli di cui all'art. 2135 del C.C., i piccoli imprenditori di cui all'art. 2083 dello stesso codice e le società semplici. Le persone fisiche, le società e i consorzi iscritti negli albi di cui alla Legge 8/8/1985, n. 443, sono annotati nella medesima sezione speciale.*

5. STRADA

Definizione

Secondo quanto disposto dall'art. 2 del D. Lgs. 30 aprile 1992, n. 285 (Nuovo codice della strada) *si definisce* ***"strada"*** *l'area ad uso pubblico destinata alla circolazione dei pedoni, dei veicoli e degli animali.*

Comma 2. *Le strade sono classificate, riguardo alle loro caratteristiche costruttive, tecniche e funzionali, nei seguenti tipi:*
A – autostrade;
B – strade extraurbane principali;
C – strade extraurbane secondarie;
D – strade urbane di scorrimento;
E – strade urbane di quartiere;
F – strade locali.

Comma 5. *Per le esigenze di carattere amministrativo e con riferimento all'uso e alle tipologie dei collegamenti svolti, le strade, come classificate ai sensi del comma 2, si distinguono in strade "statali", "regionali", "provinciali", "comunali"*, sulla base delle indicazioni contenute nella stessa norma. *Enti proprietari di dette strade sono rispettivamente lo Stato, la regione, la provincia, il comune. Per le strade destinate esclusivamente al traffico militare e denominate "strade militari", ente proprietario è considerato il comando della regione militare territoriale.*

Archivio

Comma 8. Il Ministero delle Infrastrutture e dei Trasporti (per le strade statali) e le Regioni (per le rimanenti strade) provvedono, sulla base dei criteri e con le modalità indicate dalla stessa norma, alla classificazione delle strade. *Le strade così classificate sono iscritte* ***nell'archivio nazionale delle strade*** *previsto dall'art. 226* del citato D.Lgs. 285/92

Con il richiamato art. 226 (Organizzazione degli archivi e dell'anagrafe nazionale) *presso il Ministero dell'infrastrutture e dei trasporti è istituito l'archivio nazionale delle strade, che comprende tutte le strade distinte per categoria, come indicato nell'art. 2* (D.Lgs 285/92).
Nell'archivio nazionale, per ogni strada, devono essere indicati i dati relativi allo stato tecnico e giuridico.
La raccolta dei dati avviene attraverso gli enti proprietari della strada che sono tenuti a trasmettere all'Ispettorato generale per la circolazione e la sicurezza stradale tutti i dati relativi allo stato tecnico e giuridico delle singole strade.

N.B.: Dal sito Internet del Ministero delle infrastrutture e dei trasporti - Ispettorato generale per la circolazione e la sicurezza stradale - risulta che l'Archivio nazionale delle strade è ancora in corso di costituzione.

Iscrizione e Variazioni

Comma 9. *Quando le strade non corrispondono più all'uso e alle tipologie di collegamento previste sono declassificate dal Ministero delle infrastrutture e dei trasporti e dalle regioni, secondo le rispettive competenze.*

L' Art. 13 (Norme per la costruzione e la gestione delle strade), comma 6 del D.Lgs 285/92 prevede che: *gli enti proprietari delle strade sono obbligati ad istituire e tenere aggiornati la cartografia, il catasto delle strade e le loro pertinenze secondo le modalità stabilite con apposito decreto* emanato dal Ministro delle infrastrutture e dei trasporti.

N.B.: con Decreto del Ministero dei lavori pubblici 1 giugno 2001, sono state approvate le modalità di istituzione ed aggiornamento del Catasto delle Strade, di cui al co. 6 dell'art. 13 del D.Lgs. 285/92.

5.1 STRADA COMUNALE

Definizione

Nell'art. 2, co.6 lett.D, del D.Lgs. 30 aprile 1992, n. 285 (Nuovo codice della strada) si definiscono strade comunali quelle che *congiungono il capoluogo del comune con le sue frazioni o le frazioni tra loro, ovvero congiungono il capoluogo con la stazione ferroviaria, tranviaria o automobilistica, con un aeroporto o porto marittimo, lacuale o fluviale, con interporti o nodi di scambio intermodale o con le località che sono sede di essenziali servizi interessanti la collettività comunale.* Inoltre *le strade "vicinali" sono assimilate alle strade comunali.*

Legge 24 dicembre 1954, n. 1228

(Ordinamento delle anagrafi della popolazione residente)

Art. 9 - *Il Comune provvede alla individuazione e delimitazione delle località abitate, alla suddivisione del territorio comunale in frazioni geografiche con limiti definiti in base alle condizioni antropogeografiche, ed alla esecuzione sugli adempimenti connessi.*

Art. 10 - *Il Comune provvede alla indicazione dell'onomastica stradale e della numerazione civica.*
I proprietari di fabbricati provvedono alla indicazione della numerazione interna.

D.P.R. 30 maggio 1989, n. 223

(Approvazione del nuovo regolamento anagrafico della popolazione residente)

Art. 41 (Adempimenti ecografici) – *Ogni area di circolazione deve avere una propria distinta denominazione (co. 1).*
Costituisce area di circolazione ogni spazio (piazza, piazzale, via, viale, vicolo, largo, calle e simili) del suolo pubblico o aperto al pubblico destinato alla viabilità (co. 2).
L'attribuzione dei nomi deve essere effettuata secondo le norme di cui al regio decreto legge 10/5/1923 n. 1158[2]*, convertito dalla legge 17/4/1925, n. 473, e dalla legge 23/6/1927, n. 1188*[3] *(co. 3).*
In caso di cambiamento di denominazione dell'area di circolazione deve essere indicata anche la precedente denominazione (co.4).
Nell'ambito del territorio comunale non può essere attribuita una stessa denominazione ad aree di circolazione dello stesso tipo, anche se comprese in frazioni amministrative diverse (co. 5).

Sulla base di quanto previsto dagli artt. 42 (numerazione civica) e 43 (obblighi dei proprietari dei fabbricati) le aperture che si affacciano sul fronte stradale (abitazioni, locali piano terra, passi carrabili, etc.) devono essere provvisti di numeri civici, su richiesta dei cittadini interessati.
L'obbligo della numerazione si estende anche internamente ai fabbricati per gli accessi che immettono nelle abitazioni o in ambienti destinati all'esercizio di attività professionali, commerciali e simili.
Qualora l'indicazione della numerazione interna non venga effettuata dal proprietario, vi provvede il comune.

Art. 45 (Stradario) – In ciascun comune l'ufficio preposto agli adempimenti ecografici deve curare la compilazione e l'aggiornamento dello stradario secondo le indicazioni fornite dall'Istat.

[2] *Le amministrazioni municipali, qualora intendano mutare il nome di qualcuna delle vecchie strade o piazze comunali, dovranno chiedere ed ottenere preventivamente l'approvazione del Ministero dell'istruzione pubblica per il tramite delle competenti sovraintendenze ai monumenti.*

[3] *Nessuna denominazione può essere attribuita a nuove strade e piazze pubbliche senza l'autorizzazione del Prefetto, udito il parere della regia deputazione di storia patria, o, dove questa manchi, della società storica del luogo o della regione.*

6. IL COMUNE

Definizione

Sulla base di quanto previsto dal D.Lgs. 18 agosto 2000, n. 267 (Testo unico delle leggi sull'ordinamento degli enti locali), il Comune è un ente locale che ha una propria autonomia statuaria, normativa, organizzativa e amministrativa ed è titolare di funzioni proprie e di quelle conferitegli con legge dello Stato e della Regione (artt. 2 e 3).

I comuni, come anche le province, adottano il proprio statuto con il quale sono stabilite le norme fondamentali dell'organizzazione dell'ente e regolamenti nelle materie di propria competenza (artt. 6 e 7).

Spettano al Comune tutte le funzioni amministrative che riguardano la popolazione ed il territorio comunale, dell'assetto ed utilizzazione del territorio (art. 13).

Modifiche territoriali, fusione ed istituzione di Comuni (art. 15)

A norma degli artt. 117 e 133 della Costituzione, le Regioni possono modificare le circoscrizioni territoriali dei Comuni sentite le popolazioni interessate, nelle forme previste dalla Legge regionale. Salvo i casi di fusione tra più comuni, non possono essere istituiti nuovi comuni con popolazione inferiore ai 10.000 abitanti o la cui costituzione comporti, come conseguenza, che altri comuni scendano sotto tale limite.

La denominazione delle borgate e frazioni è attribuita ai comuni ai sensi dell'art. 118 della Costituzione.

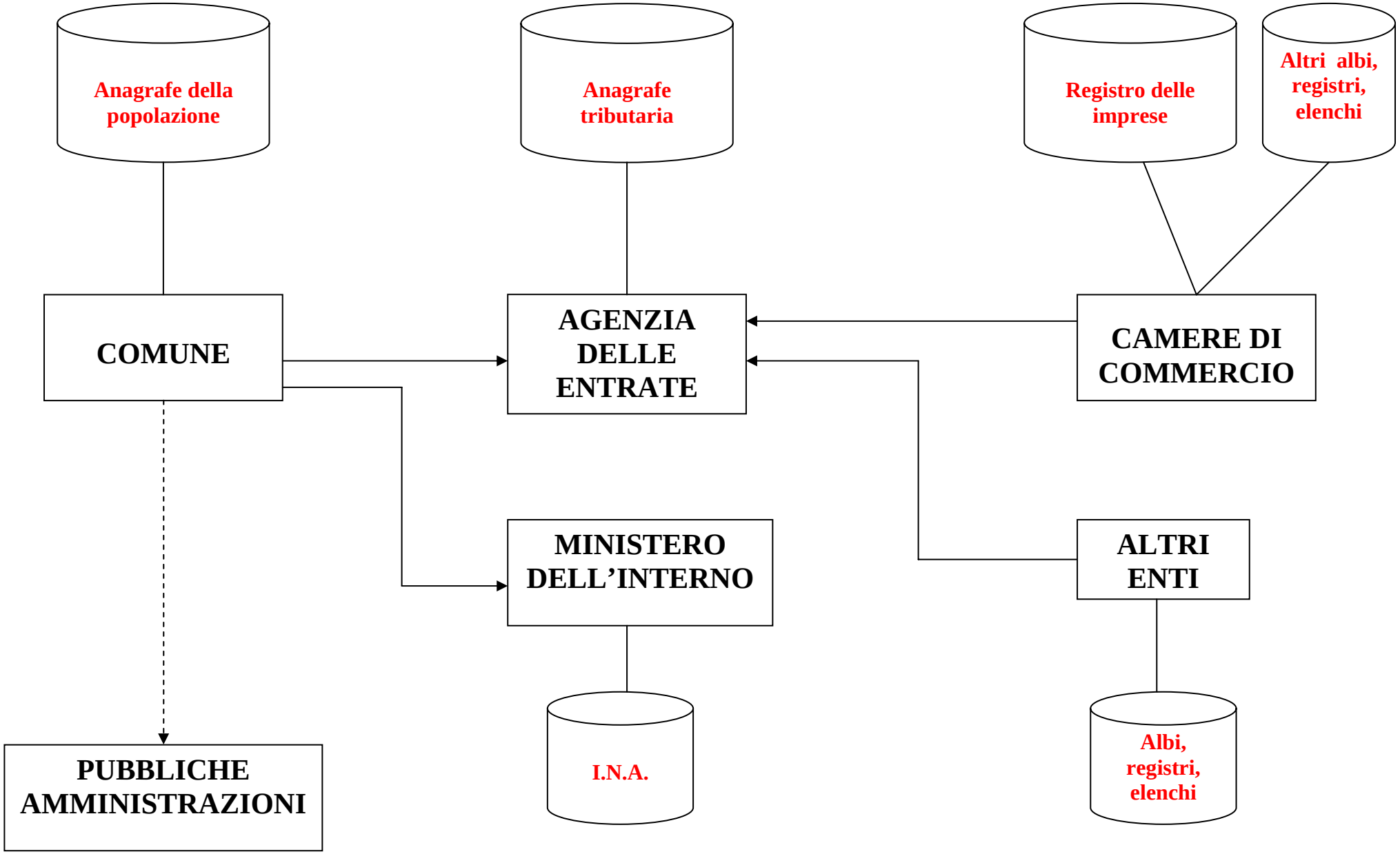
Allegato 1: Flussi informativi

L'art. 34 del D.P.R. 223 del 30 maggio 1989 prevede che *alle amministrazioni pubbliche che ne facciano motivata richiesta, per esclusivo uso di pubblica utilità, l'ufficiale di anagrafe rilascia, anche periodicamente, elenchi degli iscritti nella anagrafe della popolazione residente.*

L' art. 2 comma 1 del D.L. 15 gennaio 1993, n. 6 prevede che *i rapporti tra pubbliche amministrazioni e quelli intercorrenti tra queste e altri soggetti pubblici o privati devono essere tenuti sulla base del codice fiscale.*
Con tale norma e con il successivo D.P.C.M. 5 maggio 1994, viene quindi previsto l'inserimento del codice fiscale nelle anagrafi comunali, quale chiave identificativa dei soggetti per lo scambio dei dati. L'obbligatorietà dell'inserimento dei codici fiscali nelle anagrafi comunali è confermata dall'art. 2, comma 2, del D.M. 6 ottobre 2000, ai fini del rilascio della carta e del documento di identità elettronici.
Ai sensi dell'art. 3, comma 2, del predetto D.M. 6/10/2000, per mantenere costante nel tempo l'allineamento delle informazioni anagrafiche con i relativi codici fiscali è necessario che i comuni provvedano a inviare telematicamente all'Agenzia delle Entrate ogni variazione intervenuta, mano a mano che la stessa viene registrata nello schedario della popolazione residente.
E' inoltre previsto (circolare del Ministero dell'Interno del 5 febbraio 2001, n.1) che, ai fini del rilascio della carta di identità elettronica, che a regime verrà effettuato da parte di tutti i comuni, ogni variazione anagrafica deve essere contestualmente comunicata anche al Ministero dell'Interno, tramite il Sistema di Accesso e Interscambio Anagrafico (S.A.I.A.).

Sulla base di quanto previsto dagli artt. 6 e 7 del D.P.R. 29 settembre 1973, n. 605, ***le camere di commercio, industria, artigianato ed agricoltura devono comunicare*** (mensilmente) all'anagrafe tributaria i dati e le notizie contenuti nelle domande di iscrizione, variazione e cancellazione nei registri delle ditte, negli albi degli artigiani e negli albi registri ed elenchi istituiti per l'esercizio di attività professionali e di altre attività di lavoro autonomo.
Gli ordini professionali e gli altri enti ed uffici preposti alla tenuta di albi registri ed elenchi, devono comunicare alla anagrafe tributaria le iscrizioni, variazioni e cancellazioni.

Ai sensi dell'art. 12 del D.Lgs. 18 agosto 2000, n. 267, gli enti locali esercitano i compiti conoscitivi e informativi concernenti le loro funzioni in modo da assicurare, anche tramite sistemi informativo-statistici automatizzati, la circolazione delle conoscenze e delle informazioni tra le amministrazioni, per consentirne, quando previsto, la fruizione su tutto il territorio nazionale.


Anagrafe della popolazione
Anagrafe tributaria
Registro delle imprese
Altri albi, registri, elenchi
COMUNE
AGENZIA DELLE ENTRATE
CAMERE DI COMMERCIO
MINISTERO DELL'INTERNO
ALTRI ENTI
PUBBLICHE AMMINISTRAZIONI
I.N.A.
Albi, registri, elenchi

Appendice 3. Definizione di indicatori di qualità per dati toponomastici

1. Introduzione

La possibilità di individuare sul territorio un determinato soggetto è strettamente dipendente dalla qualità delle variabili riferite alla localizzazione, singolarmente o congiuntamente considerate.

In questo contesto per "variabili riferite alla localizzazione" si farà riferimento al seguente *insieme minimo di componenti di un indirizzo* (si veda a riguardo il documento di Crescenzi, Gargano e Boggia):

1. Provincia;
2. Comune;
3. DUG (Denominazione Urbanistica Generica);
4. Denominazione area di circolazione;
5. Numero civico

Osserviamo che per la variabile "provincia" si farà riferimento alla sua "sigla", mentre nel caso della variabile "comune" ci si riferisce alla denominazione. La variabile CAP, non essendo strettamente indispensabile ai fini del riconoscimento dell'indirizzo può svolgere la funzione di variabile ausiliaria, come sarà illustrato nel seguito.

La qualità è strettamente legata alla presenza o meno di errori nelle variabili. A livello di singole variabili i principali errori che si riscontrano sono:

(a) mancanza del valore (*completezza*);

(b) valore non ammissibile.

Nella situazione (a) si è in presenza di un *errore di completezza* (cfr. par. 2.3). L'omissione di un valore può compromettere la localizzazione di un soggetto, soprattutto se tale valore è quello del comune o della denominazione dell'area di circolazione.

Situazioni di tipo (b) si verificano quando il valore osservato per una variabile è riportato in modo errato; ad esempio, il valore esatto di una denominazione dell'area di circolazione è "Giuseppe Garibaldi" mentre il dato registrato è "Giuseppe Gariboldi".

Entrambi questi errori danno origine ad un errore di *accuratezza sintattica*, che può essere definita come:

> la vicinanza tra il *valore di una variabile riferita alla localizzazione* di uno specifico soggetto e il *valore ufficiale* della denominazione o della codifica della variabile.

Le variabili riferite alla localizzazione di un soggetto non sono tra loro indipendenti. Questo implica che la combinazione di due o più valori, che singolarmente considerati siano sintatticamente corretti, può comunque inficiare la localizzazione del soggetto stesso laddove tali valori non siano tra loro coerenti. Un esempio è rappresentato da una denominazione di un comune ammissibile (appartenente all'insieme dei comuni italiani) ma non appartenente alla provincia indicata. Situazioni di questo tipo danno origine a errori di *consistenza interna*.

Si noti che anche nel caso in cui tutte le variabili presentino valori sintatticamente corretti e tra loro coerenti, è possibile che l'insieme di tali valori non consenta di localizzare materialmente il soggetto a cui gli stessi erano riferiti. In questo caso si è in presenza di errori di *accuratezza semantica*. Poiché quest'ultimo aspetto esula dagli obiettivi definiti nel piano di dettaglio, queste particolari tipologie di errori non verranno ulteriormente analizzate.

Sulla base di tali considerazioni si è proceduto alla definizione di un insieme di indicatori di qualità riferiti a singole variabili e a loro combinazioni. Questi ultimi sono stati introdotti per tenere conto delle relazioni esistenti tra alcune variabili.

1.1 Simbologia introduttiva

Ai fini formali, conviene introdurre la simbologia di riferimento (Tabella 1).

Tab. 1 - Simbologia adottata per la definizione degli indicatori

Simbolo	**Significato**
i (i =1,…,N)	Indice che individua la singola unità presente in un dato archivio amministrativo
j (j = 1,…, 5)	Indice che individua la singola variabile dell'indirizzo
$y_{i,j}$	Variabile dicotomica che assume valore 1 se la variabile j dell'indirizzo appartenente all'unità i è sintatticamente **non** corretta, e valore 0 altrimenti.
$z_{i,j}$	Variabile dicotomica che indica la presenza o meno di un valore per la variabile j-esima dell'unità i-esima e che assume il valore 1 se l'unità i presenta un valore mancante per la variabile j, e valore 0 altrimenti.

2. Accuratezza sintattica

Come anticipato in precedenza, l'accuratezza sintattica può essere intesa come la vicinanza tra il *valore di una variabile riferita alla localizzazione* di uno specifico soggetto e il *valore ufficiale* della denominazione o della codifica della variabile. Tale definizione include sia errori di completezza che errori dovuti al fatto che un valore osservato (non mancante) sia diverso da quello ufficiale.

Ai fini dell'introduzione degli indicatori di accuratezza sintattica ipotizzeremo l'esistenza di un ipotetico *archivio di riferimento* aggiornato e contenente tutti i valori ammissibili per ciascuna delle variabili riferite alla localizzazione.

Si fa presente che sono disponibili sul mercato dei programmi automatici di riconoscimento degli indirizzi, che trattano tutte le variabili congiuntamente tenendo conto delle relazioni esistenti tra le stesse. In virtù di ciò, spesso i programmi in questione sono in grado di correggere taluni errori sintattici presenti nelle singole variabili. Per il momento l'analisi sarà limitata alla considerazione di un generico software di confronto privo di tale caratteristica.

2.1 Indici di accuratezza sintattica per singola variabile e complessivo

Per ciascuna variabile costituente l'indirizzo, è possibile calcolare cinque differenti indicatori di *Errore Sintattico* (ES_j):

$$ES_j = \frac{1}{N}\sum_{i=1}^{N} y_{i,j}\ , \qquad j=1,\ldots,5. \tag{1}$$

Un indicatore che fornisca una misura complessiva del grado di non accuratezza sintattica presente in un archivio può essere definito nel seguente modo:

$$ESG = \frac{1}{N \times 5}\sum_{i=1}^{N}\sum_{j=1}^{5} y_{i,j} \tag{2}$$

dove *ESG* sta per *Errore Sintattico Globale*.

E' possibile costruire anche un indicatore globale ponderato:

$$ESG_w = \sum_{j=1}^{5} w_j ES_j \qquad (3)$$

in cui w_j $(0 \le w_j \le 1)$ è il peso assegnato alla variabile *j*-esima. Il sistema di ponderazione deve essere tale che $\sum_{j=1}^{5} w_j = 1$.

Un criterio di definizione dei pesi potrebbe essere quello di tenere conto della rilevanza di ciascuna variabile nel riconoscimento complessivo dell'indirizzo al termine della applicazione del software di riconoscimento. In particolare, a partire da una tabella 2x2 del tipo:

Errore sintattico nella variabile *j*	Riconoscimento SI = 0	NO = 1	Tot.
NO = 0	n_{11}	n_{12}	$n_{1\bullet}$
SI = 1	n_{21}	n_{22}	$n_{2\bullet}$
Tot.	$n_{\bullet 1}$	$n_{\bullet 2}$	n

Si potrebbe stimare il *coefficiente di incertezza* (Agresti, 1990, p. 25):

$$u_j = -\frac{\sum_{r=1}^{2}\sum_{c=1}^{2} \frac{n_{rc}}{n} log\left(\frac{n \cdot n_{rc}}{n_{r\bullet} \cdot n_{\bullet c}}\right)}{\sum_{c=1}^{2} \frac{n_{\bullet c}}{n} log\left(\frac{n_{\bullet c}}{n}\right)} \qquad (4)$$

che esprime la quota di incertezza (entropia) della variabile riconoscimento spiegata dalla presenza o meno di errore nella variabile in esame (*j*). Esso può variare tra 0 ed 1 $(0 \le u_j \le 1)$. Una volta calcolato tale coefficiente per tutte le variabili in esame, il peso di ciascuna da utilizzare ai fini del calcolo della (3) sarà dato da:

$$w_j = \frac{u_j}{\sum_{j=1}^{5} u_j}, \quad j=1,\ldots,5. \qquad (5)$$

Si noti che in alternativa al coefficiente di incertezza si può far riferimento anche ad altre misure di associazione tra variabili categoriali (cfr. Agresti, 1990).

2.2 Accuratezza sintattica per sottogruppi di variabili

Gli indici di accuratezza sintattica per ciascuna singola variabile, così come quello complessivo, non permettono di verificare la presenza o meno di particolari configurazioni (*pattern*) di errori per cui, ad esempio, in un certo archivio la

presenza di un errore nella denominazione dell'area di circolazione potrebbe presentare un elevato grado di associazione con errori nel numero civico. Una tale informazione può risultare utile per revisionare l'archivio in questione, perché può mettere in condizione di individuare le cause che hanno prodotto il particolare *pattern* di errori e quindi rimuoverle.

Operativamente, per individuare eventuali pattern di errori sintattici risulta conveniente costruire una variabile (stringa) composta di tante posizioni quante sono le variabili che compongono l'informazione toponomastica. In ogni posizione della stringa vi sarà un 1 o uno 0 a seconda che la variabile, cui la posizione si riferisce, presenti o meno un valore sintatticamente non corretto. Così, ad esempio, $\phi_i = 00000$ indica che l'unità *i*-esima presenta dei valori sintatticamente corretti per tutte le variabili in questione. Al contrario, $\phi_i = 00010$ indica che l'unità *i*-esima presenta valori esatti per tutte le variabili ad eccezione di quella in posizione 4, ovvero la denominazione dell'area di circolazione ($j = 4$).

La variabile ϕ_i può presentare 2^5 possibili configurazioni. L'individuazione di particolari *pattern* di valori sintatticamente errati potrà essere condotta contando nell'archivio in questione il numero di occorrenze (assolute e relative) di ogni possibile configurazione.

Si fa presente che un'analisi di questo tipo può anche riguardare un sotto-insieme significativo di variabili come, ad esempio, quello relativo a comune, DUG e denominazione dell'area di circolazione. In tal caso, sarà conveniente definire una sotto-stringa della precedente di sole tre posizioni. Così, ${}_{234}\phi_i = 010$ indica una situazione in cui la DUG è errata ($j = 3$) mentre risultano corretti sia il comune ($j = 2$) che la denominazione dell'area di circolazione ($j = 4$). Anche questa stringa di dimensioni ridotte dovrà essere analizzata secondo quanto illustrato per quella completa.

2.3 Completezza

Come illustrato in precedenza, la completezza può essere vista come una componente della accuratezza sintattica. Laddove tale errore abbia una forte incidenza sull'archivio in esame, può risultare conveniente analizzarlo separatamente.

E' bene tener presente che in questo contesto si fa riferimento alla seguente definizione di completezza:

> una componente dell'indirizzo, cioè una variabile riferita alla localizzazione di un soggetto, si definisce *completa* se presenta un valore (prescindendo dalla effettiva correttezza del valore osservato).

In pratica, la completezza viene valutata a livello di informazione elementare, e viene rilevata semplicemente in termini di presenza o assenza di un valore per una

delle variabili precedentemente definite. La completezza, tuttavia, può essere riferita anche ad altri aspetti. In particolare, a livello dell'intero archivio, il concetto di completezza corrisponde al concetto di *copertura* dell'archivio rispetto alla popolazione teorica di riferimento, e si riferisce alla capacità dell'archivio di includere tutte e sole le unità ad essa appartenenti. Questo aspetto esula dagli obiettivi del progetto definiti nel piano di dettaglio.

Da un punto di vista pratico, la completezza può essere valutata sia in relazione alla singola variabile sia per sottogruppi rilevanti delle variabili precedentemente introdotte.

2.4 Completezza per le singole variabili

In base alla notazione introdotta nel paragrafo 2, la *frazione di valori mancanti* per la variabile provincia ($j=1$) è:

$$M_1 = \frac{1}{N}\sum_{i=1}^{N} z_{i,1} \tag{6}$$

di conseguenza,

$$C_1 = 1 - M_1 = 1 - \frac{1}{N}\sum_{i=1}^{N} z_{i,1} \tag{7}$$

rappresenta la *frazione di valori presenti* per la variabile provincia.

Più in generale

$$C_j = 1 - M_j = 1 - \frac{1}{N}\sum_{i=1}^{N} z_{i,j} \tag{8}$$

fornisce la frazione di valori presenti per la variabile *j* (*j*=1,…,5).

Dagli indici di completezza definiti per le singole variabili è possibile ricavarne uno complessivo, definito come la media degli indici di completezza per le singole variabili

$$\overline{C} = \frac{1}{5}\sum_{j=1}^{5} C_j = 1 - \frac{1}{5 \cdot N}\sum_{j=1}^{5}\sum_{i=1}^{N} z_{i,j} \ . \tag{9}$$

Si noti che tale indice, per come è costruito, assegna uguale importanza a ciascuna delle cinque variabili che definiscono una informazione toponomastica. Laddove si voglia tener conto di una diversa importanza delle variabili è opportuno ricorrere ad una media ponderata dei singoli indici:

$$\overline{C}_w = \sum_{j=1}^{5} w_j C_j = 1 - \sum_{j=1}^{5} w_j M_j \tag{10}$$

in cui w_j è il peso assegnato alla variabile *j*-esima tale che $0 \le w_j \le 1$ e $\sum_{j=1}^{5} w_j = 1$.

Il sistema dei pesi può essere definito sulla base di quanto introdotto nel par. 2.1 con l'accortezza di sostituire la variabile presenza-assenza di errore sintattico con quella relativa alla presenza-assenza di errore di completezza.

2.5 Completezza per sottogruppi di variabili

Analogamente a quanto introdotto per l'errore di accuratezza sintattica, può risultare utile costruire una variabile (stringa) composta di tante posizioni quante sono le variabili che compongono l'informazione toponomastica. In ogni posizione della stringa vi sarà un 1 o uno 0 a seconda che la variabile, cui la posizione si riferisce, presenti o meno un valore mancante. Così, ad esempio, $\psi_i = 00000$ indica che l'unità *i*-esima presenta dei valori per tutte le variabili in questione (tutti i campi sono pieni). Al contrario, $\psi_i = 00001$ indica che l'unità *i*-esima presenta dei valori per tutte le variabili ad eccezione della variabile cui si riferisce la posizione 5, ovvero il numero civico ($j = 5$).

La variabile ψ_i può presentare 2^5 possibili configurazioni. L'individuazione di particolari *pattern* di valori mancanti potrà essere condotta contando nell'archivio in questione il numero di occorrenze (assolute e relative) di ogni possibile configurazione.

L'analisi può essere condotta anche relativamente a sotto-insiemi significativi di variabili. come, ad esempio, quello relativo a DUG e denominazione dell'area di circolazione. In tal caso, si dovrà definire una sotto-stringa della precedente di sole tre posizioni. Così, ${}_{34}\phi_i = 10$ indica una situazione in cui la DUG è mancante ($j = 3$) mentre risultano presente la denominazione dell'area di circolazione ($j = 4$). Anche questa stringa di dimensioni ridotte dovrà essere analizzata secondo quanto illustrato per quella completa.

3. Consistenza interna

La *consistenza interna* fa riferimento alle relazioni esistenti tra i valori delle variabili riferite alla localizzazione di un soggetto. Un indirizzo si definisce consistente se i valori delle variabili riferite alla localizzazione che lo compongono sono tra loro coerenti. La consistenza interna di un indirizzo può essere valutata solo se i valori delle singole variabili sono tutti presenti e sintatticamente corretti.

Le relazioni esistenti tra le diverse componenti di un indirizzo sono essenzialmente di natura gerarchica. Ai fini della valutazione della consistenza interna rivestono particolare importanza solo alcune delle possibili relazioni gerarchiche. La loro individuazione passa attraverso la suddivisione delle variabili in esame in due categorie (cfr. doc. Crescenzi, Gargano e Boggia):

- variabili che identificano la *zona di territorio*:
    a. Provincia (sigla);
    b. Comune (denominazione);
- variabili che identificano il *punto sul territorio*:

a. DUG;
b. Denominazione area di circolazione;
c. Numero civico.

Per ciò che riguarda la zona di territorio si avrà un errore di consistenza interna laddove il valore del comune, pur essendo ammissibile (il comune esiste), non è coerente con quello della provincia, cioè all'interno della provincia in questione non esiste alcun comune con tale denominazione.

Purtroppo, in situazioni di incoerenza spesso non si è in grado di identificare quale valore è errato. Infatti esistono comuni con uguale denominazione appartenenti a diverse province. E' tuttavia evidente che se esiste un solo comune che abbia una data denominazione vi sono elevate probabilità che l'errore sia nella variabile provincia.

Per quanto concerne le variabili che identificano il punto sul territorio, una volta che sia stata identificata univocamente la zona di territorio, una prima possibile situazione di incoerenza si può verificare tra DUG e denominazione dell'area di circolazione. Ad esempio, in un dato comune può non esistere "Piazza Giuseppe Garibaldi", ma soltanto "Via Giuseppe Garibaldi".

Un ulteriore caso di incoerenza si può verificare se il valore osservato per il numero civico è incoerente con i civici presenti in una data area di circolazione.

3.1 Uso di variabili ausiliarie per l'individuazione dell'errore

In situazioni di incertezza ai fini dell'individuazione dell'errore può essere utile disporre di variabili ausiliarie, come ad esempio il CAP e/o la località. Attualmente il CAP è una stringa di 5 caratteri in cui i primi due possono essere usati per identificare la provincia (CAP generico) e i restanti tre per il comune (vi è infatti il problema delle province nuove e dei comuni con lo stesso CAP). Ad esempio, laddove la provincia fosse mancante e la denominazione del comune fosse ammissibile ma non univoca il CAP può svolgere il ruolo di variabile chiave per la corretta individuazione del comune. Analogamente, il CAP può essere anche utilizzato per risolvere situazioni di errore nella denominazione del comune che hanno comportato il mancato riconoscimento da parte del software dell'intero indirizzo.

4. Accuratezza sintattica e riconoscimento dell'indirizzo

Partendo dal presupposto che i valori osservati per le variabili che compongono un indirizzo hanno lo scopo di permettere la localizzazione di un soggetto sul territorio, è evidente che la valutazione della qualità delle informazioni toponomastiche presenti in un dato archivio deve essere necessariamente legata a tale obiettivo. In tal senso appare più importante valutare la qualità dell'indirizzo nel suo complesso, tenendo quindi conto delle relazioni intercorrenti tra le variabili, piuttosto che quella delle singole componenti.

Seguendo questa impostazione, l'accuratezza dovrebbe essere valutata a livello dell'intero indirizzo. Limitando l'attenzione all'accuratezza sintattica, un indirizzo si potrà dire *sintatticamente accurato* se è possibile stabilire una corrispondenza univoca con uno degli indirizzi riconosciuti come ufficiali e presenti sul territorio nazionale.

In tal senso, la presenza di:

- errori di completezza in una o più variabili,
- errori di accuratezza sintattica in una o più variabili (singolarmente considerate),
- errori di consistenza interna,

influisce direttamente sull'accuratezza sintattica dell'intero indirizzo. Ciò è vero sia che gli errori si presentino singolarmente che in combinazione tra loro.

Nella pratica, per valutare l'accuratezza sintattica dell'indirizzo non si può prescindere dalla considerazione dello strumento che si utilizza per mettere in relazione gli indirizzi dell'archivio in esame con quelli dell'archivio di riferimento. In particolare, è necessario tener conto delle caratteristiche di funzionamento dell'algoritmo di riconoscimento degli indirizzo che si sta utilizzando. Ad esempio, un algoritmo può avere caratteristiche tali da rendere ininfluenti ai fini del riconoscimento alcuni errori di completezza o di accuratezza sintattica in singole variabili. Casi tipici sono l'assenza della provincia in casi in cui il comune presenta denominazione unica sul territorio; ovvero piccoli errori di trascrizione in una denominazione ("Giribaldi" anziché "Garibaldi").

In tal senso, se si fa riferimento ad una procedura automatica di riconoscimento, un ovvio indicatore generico dell'accuratezza sintattica dell'archivio può essere definito anche come la frazione di indirizzi dell'archivio stesso riconosciuti dal software:

$$R_0 = 1 - \frac{1}{N}\sum_{i=1}^{N} r_i \qquad (11)$$

dove

$$r_i = \begin{cases} 0, & \text{se l'}i\text{-esimo indirizzo è stato riconosciuto;} \\ 1, & \text{altrimenti} \end{cases}$$

Per valutare la relazione tra accuratezza sintattica di un indirizzo e il riconoscimento o meno delle stesso da parte dell'apposito software può risultare conveniente "allungare" la variabile stringa, ϕ_i, introdotta nel par. 2.2, accodando ad essa la variabile indicatrice r_i. La nuova stringa, denotata con ω_i, avrà quindi 6 posizioni. Il punto di partenza per qualsiasi analisi sarà quindi costituito dalla distribuzione di frequenza di tale variabile.

Nell'ambito di tale distribuzione di frequenza è possibile individuare tre situazioni che forniscono indicazioni di sintesi sulla relazione tra accuratezza sintattica e riconoscimento:

a) "indirizzo corretto": l'indirizzo è sintatticamente accurato ($\phi_i = 00000$) e viene riconosciuto dal software ($r_i = 0$). La nuova stringa è $\omega_i = 000000$.

b) violazione "debole" dell'accuratezza sintattica dell'indirizzo: vi è almeno un 1 in una delle posizioni di ϕ_i ma l'indirizzo viene comunque riconosciuto dal software ($r_i = 0$). In tal caso l'errore di accuratezza sintattica risulta ininfluente ai fini del riconoscimento.

c) Violazione "grave" dell'accuratezza sintattica dell'indirizzo: vi è almeno un 1 in una delle posizioni di ϕ_i e l'indirizzo non viene riconosciuto dal software ($r_i = 1$). In tal caso l'errore (o gli errori) di accuratezza sintattica influenza i risultati del riconoscimento.

La frequenza relativa di ciascuna delle tre situazioni appena elencate nell'ambito dell'intero archivio rappresenta un indicatore di carattere generale della qualità di un archivio. In particolare, la frequenza relativi degli "indirizzi corretti", ossia quali per i quali risulta $\omega_i = 000000$, può essere denotato come *tasso di accuratezza sintattico.*

Si fa presente, che l'elenco degli indicatori presentati non è da ritenersi esaustivo, numerosi altri se ne possono costruire tenendo presente il funzionamento dell'algoritmo di riconoscimento e la struttura gerarchica delle relazioni tra variabili costituenti l'indirizzo.

Bibliografia

Agresti, A. (1990) *Categorical Data Analysis*. Wiley, New York.

Crescenzi F., Gargano O. e Boggia A. (2003) *Proposta di Standard dei Dati Toponomastici.* Documento prodotto nell'ambito del progetto AIPA-ISTAT, Giugno 2003.