Controllo qualità dei dati e interoperabilità in aggiornamento

Alberto Belussi (*), Mauro Negri(**)

(*) Università di Verona, Dipartimento di Informatica, Verona, Italy, alberto.belussi@univr.it
(**) Politecnico di Milano, Dipartimento DEI, Piazza L.Da Vinci 32, 20132 Milano, Italy, negri@elet.polimi.it

Riassunto
L'interoperabilità dei dati in una SDI di database topografici nei quali si scambiano dati o gli aggiornamenti degli stessi impone di avere una architettura che garantisca l'interscambio dei dati tra un insieme ampio ed eterogeneo di soggetti. Questo lavoro definisce le caratteristiche minime che i soggetti locali della SDI devono rispettare per garantire l'interoperabilità garantendo la massima indipendenza dalle tecnologie utilizzate a livello locale.

Abstract
This papers deals with interoperability of data in an integrated SDI in which several actors provide data and their updates for a portion of the whole territory and the SDI must guarantee their integration and harmonization. In particular, several crucial aspects for the realization of a real interoperability of data and of the updates are listed and analyzed.

Introduzione
Il lavoro svolto dall'IntesaGIS prima e dal comitato per le regole tecniche sui dati territoriali delle PA poi, ha permesso la definizione del contenuto del database topografico (DBT) da realizzare a livello regionale e nazionale (National Core). Diversi progetti regionali, alcuni dei quali supportati dal CISIS stanno sperimentando la produzione del DBT o lo hanno parzialmente già realizzato. La realizzazione dei DBT costituisce solo il primo passo per la costruzione di una SDI nazionale che coinvolge diversi attori dal livello locale a quello nazionale; i DBT locali dei comuni o dei centri servizi gestiscono i dati e forniscono le sintesi necessarie a livello regionale e nazionale. Per ottenere questo risultato è necessario che il sistema complessivo garantisca l'interoperabilità dei dati tra i diversi soggetti coinvolti e l'aggiornamento dei dati per garantire il mantenimento del sistema; l'aggiornamento dovrebbe essere eseguito da coloro che sono responsabili del dato e propagato poi agli altri soggetti, può essere periodico e interessare larghe porzioni di territorio o essere frequente ed essere apportato non appena si verifica e richiede l'integrazione e l'armonizzazione dei dati tra soggetti confinanti. L'interscambio ad esempio tra comuni e regione è complesso poiché deve garantire la massima indipendenza dei singoli attori nella scelta delle proprie tecnologie e d'altra parte deve permettere di integrare e armonizzare i dati locali per la coerenza del DBT regionale o nazionale garantendo una gestione centralizzata dei grandi oggetti condivisi. In questo contesto diventa fondamentale avere un modo unico e condiviso per il trasferimento e l'interpretazione dei dati oggetto di interscambio. Normalmente si confonde spesso l'interoperabilità con la definizione del solo formato fisico di interscambio, ma sebbene importante è solo la definizione di un aspetto sintattico dell'interoperabilità; infatti non appena definito il formato diventa fondamentale la capacità di associare la stessa semantica ai dati condivisi da parte dei vari soggetti. Per questo motivo prendendo spunto anche dalle esperienze nella produzione di DBT, in particolare di quello della Regione Lombardia, si identificheranno le caratteristiche che definiscono l'interoperabilità dei dati utilizzando come criterio di riferimento quello di definire tali caratteristiche in modo che siano indipendenti da una specifica tecnologia (GIS, DBMS,…). Si tenga presente che tali caratteristiche saranno da verificare e arricchite nel processo di realizzazione delle SDI e che la loro definizione è

presupposto per la definizione dei controlli che ne garantiscano la qualità attraverso procedure anch'esse indipendenti dalla specifica tecnologia. Nel seguito si descrivono tali caratteristiche definendo lo stato dell'arte e future possibili evoluzioni.

Modello dei dati condiviso
I dati di una SDI, indipendentemente dalla materializzazione o meno dei dati complessivi, sono memorizzati in sistemi diversi, con rappresentazioni differenti e trasferiti con formati talvolta non necessariamente identici. Diventa pertanto importante definire un modello dei dati (concettuale) in grado di descrivere le proprietà rilevanti dei dati astraendo dai dettagli tecnici specifici delle diverse tecnologie coinvolte. Il modello deve prevedere la definizione delle caratteristiche formali dei dati e quindi controllabili e di quelle informali e quindi non verificabili automaticamente ma comunque utili per la comprensione dei dati. Il modello definisce il concetto di classe, attributi descrittivi, identificatori, componenti spaziali e loro attributi e il testo informale può descrivere come identificare un oggetto della classe (ad esempio, come un'area stradale si spezzi in un incrocio). A livello internazionale è stato adottato l'UML come modello standard per questo fine dall'OGC e dall'ISO TC211 ed è stato utilizzato da INSPIRE per la definizione dei propri schemi condivisi e pertanto può essere considerato il modello di riferimento.

Modello della geometria
L'OGC e l'ISO TC211 si sono concentrati sulla standardizzazione di un modello geometrico da utilizzare per le componenti spaziali dei dati (inesistente nell'UML) che fosse indipendente dalla specifica realizzazione tecnologica. E' stato definito lo standard 19107 (iso19107) che descrive i modelli geometrici e topologici in un'ottica "full 3D" che è da considerare il riferimento generale. Inoltre è stato definito un profilo del modello generale allo scopo di definire un modello a cui ogni tecnologia si debba conformare chiamato Simple Feature Model (SFM)(iso19125).Questo modello definisce una gerarchia di tipi geometrici (punti, curve, superfici e collezioni di questi tipi elementari) descrivendone le caratteristiche nello spazio euclideo (ad es., che una superficie semplice è una porzione di piano fortemente connessa e confinata da una curva semplice e chiusa). Definisce inoltre le operazioni disponibili sui diversi tipi tra le quali le relazioni topologiche applicabili (ad es., disjoint, touch,…).
SFM è un modello 2D, sebbene si stia lavorando ad una sua estensione parziale verso il 3D, e questo ne limita le potenzialità descrittive. A livello italiano è stata proposta una varianet del SFM incorporata nel modello GeoUML (Belussi, 2010) proposto da SpatialDBgroup del Politecnico di Milano e adottata dal comitato per le regole tecniche sui dati territoriali delle PA per i DBT. Il geoUML si basa sull'UML e adotta il modello SFM della geometria estendendolo ai punti e curve 3D e alle superfici B3D ossia superfici che hanno la frontiera definita in 3D e la superficie (usata per le relazioni topologiche) ottenuta dalla proiezione della frontiera nel piano 2D.

I vincoli spaziali e metrici
Le componenti spaziali dei dati di uno stesso territorio interagiscono tra di loro spazialmente in base a vincoli più o meno forti (ad es., un marciapiede può essere al più adiacente all'impronta al suolo di un edificio). L'OGC e l'ISO TC 211 hanno riconosciuto la necessità di descrivere tali vincoli come come parte aggiuntiva alla descrizione concettuale, ma si sono limitati ad indicare il linguaggio Object Constraint Language (OCL) dell'UML come lo strumento formale per la definizione dei vincoli. Non sono classificati i tipi di vincoli possibili e non viene specificato come usare il linguaggio. Lo stesso approccio è stato adottato anche da INSPIRE. L'OCL è un linguaggio formale di tipo generale e non orientato alla geometria che descrive un vincolo attraverso un'espressione logica. Il linguaggio è tuttavia difficile da scrivere e da leggere e ciò costituisce un grande ostacolo al suo uso da parte di personale non esperto; inoltre la sua generalità permette di descrivere uno stesso vincolo attraverso diverse formulazioni logiche creando un ostacolo alla comprensione e condivisione dei vincoli da parte di utenti diversi. Si aggiunga a questo l'inesistenza di traduttori delle formule OCL in programmi di controllo dei vincoli che rende la

gestione dei vincoli sui dati di fatto impossibile e rendendo i vincoli stessi puro testo descrittivo di scarso effetto pratico. Nell'ambito del progetto del DBT della Regione Lombardia si è invece dimostrato di cruciale importanza avere la possibilità di controllare tali vincoli. Un esempio di possibile evoluzione è dato dal GeoUML (Belussi, 2010) che preclassifica un insieme standard di vincoli spaziali basati sulle relazioni topologiche, la cui semantica è definita formalmente in OCL per conformità all'approccio ISO. La condivisione di un insieme di vincoli ha facilitato la definizione e la leggibilità dei vincoli e inoltre ciò ha permesso la predisposizione del GeoUML validator, ossia di uno strumento automatico per la verifica di tutti i vincoli definiti sui dati, ma in particolare di quelli spaziali. Infine la standardizzazione dei vincoli ha agevolato anche il riconoscimento, e quindi la correzione, degli errori rilevati dal GeoUMLvalidator.
In Figura 1 si mostrano la definizione informale INSPIRE di un vincolo spaziale sulla rete di trasporto, la descrizione in GeoUML e la formulazione in OCL.

Vincolo INSPIRE
In a Transport Networks data set which contains nodes, these nodes shall only be present where Transport Links connect or end

Vincolo GeoUML
TransportNode.Geometry (**TC**) *exists* (InNetwork = TransportNode.InNetwork) TransportLink.CenterlineGeometry

Possibile espressione OCL

```
context TransportNode
inv:TransportLink.allInstances ->
    exists(a: TransportLink |
      self.InNetwork = a.InNetwork and
      self.Geometry.check(TC,
        a.CenterlineGeometry))
```

Figura 1 – La formulazione di un vincolo spaziale.

Modelli implementativi

Le precedenti caratteristiche hanno permesso la condivisione delle caratteristiche concettuali dei dati tra utenti indipendentemente dalle tecnologie adottate. Lo schema concettuale che raccoglie dati e vincoli deve poi essere tradotto su uno specifico formato di interscambio. I formati che possono essere considerati per il trasferimento dei dati sono il gml e lo shape file dove il primo è utilizzato in INSPIRE mentre il secondo può essere considerato il formato di riferimento in attesa che il gml si consolidi. Definito il formato è pertanto necessario definire un insieme di regole generali (chiamate modello implementativo) per la traduzione degli elementi concettuali nelle strutture dati dello specifico formato di interscambio (ad es., modello implementativo per shape), evitando regole ad hoc per ogni specifico schema concettuale. Avere regole generali e quindi condivise semplifica l'interpretazione dei contenuti dei file e inoltre il disaccoppiamento concettuale/fisico permette di unificare in un solo schema gli aspetti concettuali e di avere diverse rappresentazioni fisiche dello stesso schema in funzione delle tecnologie adottate.
Questo approccio permette di:

- stabilire un preciso e semplice collegamento tra concetti e loro realizzazione fisica;
- mantenere un legame tra diverse realizzazioni fisiche della stessa struttura concettuale dei dati, favorendo anche la realizzazione di convertitori di formato ad esempio utili per convertire il contenuto del proprio database nel formato di trasferimento;
- realizzare strumenti automatici di controllo "platform independent" in grado di riconoscere le strutture fisiche dei dati, di verificare sui dati i vincoli spaziali e infine di restituire una diagnostica espressa in termini concettuali e fisici.

Il modello implementativo traduce i concetti in strutture dati fisiche, ma non traduce i vincoli che rimangono descritti nella struttura concettuale, tuttavia il modello implementativo e la conoscenza

dello schema concettuale e delle strutture fisiche permette di realizzare procedure di controllo automatiche e non ad hoc (come ad esempio il GeoUML validator).
Si noti che INSPIRE non distingue il livello concettuale da quello fisico e la definizione UML degli schemi è in realtà orientata alla sola descrizione del contenuto dei file gml sottostanti. Ciò complica la comprensione degli aspetti concettualmente rilevanti e la realizzazione di procedure di controllo di qualità di validità generale.

La rappresentazione finita dei dati
La semantica del modello geometrico e delle sue operazioni è definita supponendo di essere nel modello matematico della precisione infinita delle coordinate dei dati. Invece le coordinate sono rappresentate sul calcolatore con un modello a precisione finita e ciò implica che sia l'intervallo delle coordinate rappresentabili sia finito, ma anche che la maggior parte delle coordinate reali esprimibili in tale intervallo non siano rappresentabili. Il primo problema in pratica non esiste in quanto considerando le coordinate dell'Italia nel sistema di riferimento UTM32/WGS84 ricadono tutte nell'intervallo rappresentabile, mentre il secondo problema viene risolto arrotondando la coordinata ad una di quelle rappresentabili sul calcolatore. A complicare questo problema abbiamo che esistono vari modi di rappresentare e manipolare i dati con conseguenze diverse e complicate sulla determinazione dell'arrotondamento generato. Questo problema non è rilevante fintanto che si lavora in un solo sistema e quindi meno sentito in fase di produzione di primo impianto dei DBT, ma diventerà più grave con l'interoperabilità a causa della maggior presenza di sistemi diversi e dell'interscambio. In questa sezione si cercherà di presentare il problema e di identificare un possibile percorso euristico per la determinazione di parametri da sperimentare al fine di ridurre o annullare il problema. I sistemi adottano diverse tecniche per la rappresentazione dei dati, tuttavia quasi tutti i programmi che elaborano i dati si basano sulla rappresentazione standard floating point a 64 bit IEEE754- 1985 (nel seguito chiamata FP64) che corrisponde normalmente al concetto di numero reale in doppia precisione. Nella figura 2 sono visualizzate due curve nel mondo della precisione infinita in relazione di touch (a sinistra) e di disjoint (a destra).

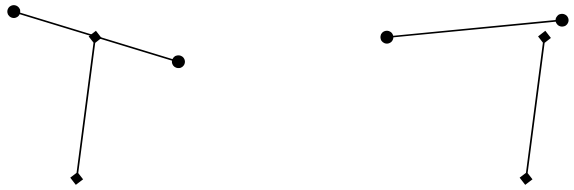

Figura 2 – Curve nello spazio euclideo.

Consideriamo ora cosa accade rappresentando le coordinate in FP64. L'FP64 descrive numeri reali in virgola mobile e distribuisce i 64 bit automaticamente per la descrizione delle parti intera e frazionaria del numero; in particolare, descrive prima la parte intera e poi dedica i bit rimanenti per la parte frazionaria e ciò significa che numeri più grandi lasciano meno bit per la parte frazionaria, comportando quindi approssimazioni maggiori delle coordinate. Considerando le coordinate degli estremi del *bounding box* dell'Italia in UTM32/WGS84 (coordinate tra 300.000 e 5.200.000) si determina che la precisione massima disponibile per la parte frazionaria delle coordinate è dell'ordine del 10^{-9} (si sottintende l'unità di misura che è il metro) in modo tale da garantire una corrispondenza biunivoca tra le coordinate decimali e le corrispondenti rappresentazioni binarie. Lo spostamento delle coordinate per l'approssimazione è dello stesso ordine di grandezza. Le coordinate sono poi manipolate dagli algoritmi geometrici (ad es., quelli di verifica delle relazioni topologiche) e ogni operazione eseguita sui dati produce a sua volta un arrotondamento nei calcoli dell'ordine di 10^{-9} metri. Algoritmi per operazioni diverse o algoritmi diversi per la stessa operazione agiscono quindi come moltiplicatori di tale approssimazione non quantificabili in modo

deterministico. Si tratta di spostamenti applicativamente non significativi, ma che possono cambiare la valutazione di un vincolo spaziale; ad esempio le geometrie precise di Figura 1 (curve continue in Figura 3) possono subire uno spostamento dei vertici e del potenziale punto di intersezione tanto da non escludere, l'inversione delle relazioni spaziali tra le geometrie (curve tratteggiate in Figura 3).

Figura 3 – Lo spostamento prodotto dalla rappresentazione finita.

I sistemi GIS cercano di ovviare al problema attraverso il concetto di tolleranza che dovrebbe annullare gli effetti di questi piccoli spostamenti; questa soluzione è utile nei sistemi locali per intercettare e correggere topologicamente i dati, ma non è praticabile per il sistema di interoperabilità per i seguenti motivi:

- sebbene la definizione di tolleranza sia intuitiva non esiste una formalizzazione di tale concetto e degli algoritmi e infatti gli standard ISOTC211 e OGC lasciano questo aspetto *implementation dependent*;
- i sistemi che si basano sulla tolleranza non specificano come gestirla e quindi non si garantisce che l'applicazione di un'operazione basata sulla tolleranza dia lo stesso risultato in sistemi differenti.

Per questi motivi si ritiene che il sistema di interoperabilità debba operare sui dati in modo preciso e definire invece dei vincoli sui dati per ovviare ai problemi dovuti alla rappresentazione finita dei dati. In particolare, si potrebbero imporre due vincoli sui dati:

- due geometrie che hanno punti di intersezione richiesti dalla relazione topologica precalcolano questi punti traducendoli in espliciti vertici coincidenti nelle geometrie (ad esempio generando un vertice nella curva orizzontale della precedente figura 3 a sinistra e imponendo che abbia lo stessa coordinata della curva verticale) in modo che qualsiasi perturbazione preservi l'uguaglianza delle coordinate;
- due geometrie che devono rimanere separate garantiscano che tra due punti o tra un punto/vertice e un segmento esista sempre una distanza minima Dm tale da mantenere le relazioni presenti nel piano euclideo di Figura 2.

La prima condizione è più semplice perché tende a imporre di spezzare tutto ciò che interseca, mentre la seconda impone di valutare Dm. Una soluzione euristica può stabilire che l'approssimazione dovuta ad un generico algoritmo non possa superare le 10/100 volte l'approssimazione della singola operazione implicando quindi che $Dm \geq 10^{-6}$.

Si consideri ora all'estremo opposto l'accuratezza del rilievo che è dell'ordine del 10^{-2} e la possibilità di rappresentare anche geometrie fittizie nella giusta dimensionalità (ad esempio, una parete verticale rappresentata come una piccolissima superficie ottenuta inclinando leggermente la parete stessa o una superficie degenerata a curva perché sotto la soglia di cattura); si può ipotizzare che queste geometrie siano descritte generando delle superficie in cui vertici e curve distino meno dell'accuratezza del rilievo, ad esempio millimetri (10^{-3}) quando l'accuratezza è al centimetro. Questa ipotesi determina che la risoluzione di riferimento non possa essere inferiore al millimetro necessario per le geometrie fittizie, ma d'altra parte impone un ulteriore vincolo a Dm che quindi viene ristretta all'intervallo

$$10^{-4} \leq Dm \leq 10^{-6}$$

per non ritenere troppo vicine le geometrie fittizie da un lato e da inglobare gli errori di computazione dall'altra. Rimane infine da valutare l'impatto delle perturbazione delle geometrie dovuto all'introduzione di nuove intersezioni. Ogni volta che una geometria viene spezzata per introdurre un vertice a causa di un'intersezione con un'altra geometria si produce uno spostamento della geometria originaria (ad esempio, un segmento di retta diventa una spezzata di due segmenti non più collineari); tale spostamento è tanto più elevato tanto più la precisione è bassa (spostamenti di millimetri per precisioni al millimetro). La risoluzione di riferimento dei dati deve quindi essere più fine di quelle delle geometrie rilevate o fittizie e più alta di quella del FP64 per incorporare gli errori computazionali, ossia $10^{-4} \leq Rrif \leq 10^{-6}$.
Nelle sperimentazioni regionali effettuate nell'ambito dei progetti del CISIS si è scelto di adottare una risoluzione di riferimento $Rrif = 10^{-5}$ e una distanza minima $Dm = 10^{-4}$. Valori più piccoli non sono stati considerati anche per non complicare la rilevazione degli eventuali errori individuati nei dati negli usuali strumenti GIS.
Ciò ha un impatto anche sul controllo di qualità dei dati che deve quindi eseguire i calcoli con una risoluzione più fine di quella adottata nel sistema in modo da non introdurre ulteriori perturbazioni, che non deve inoltre adottare meccanismi basati sulla tolleranza e tantomeno deve correggere topologicamente i dati prima dei controlli.

Indici di accettabilità e metainformazione
La definizione condivisa delle proprietà dei dati permette il controllo della qualità dei dati. Le esperienze condotte in rilievo dimostrano che il controllo esaustivo dei vincoli definiti sui dati può trovare molti errori: errori sistematici o casuali che saranno corretti nel processo di consegna o aggiornamento. Diventa quindi importante arricchire i dati con la metainformazione di istanza che colleghi ogni errore all'oggetto considerato in modo che il soggetto gestore possa valutare la qualità dei singoli oggetti ricevuti. Inoltre è necessario definire nel sistema di interoperabilità degli indici di accettabilità che indichino dei livelli di errore che possono rendere complessivamente o parzialmente inaccettabile un insieme dei dati e questo è particolarmente importante nel caso di trasmissione di aggiornamenti che dovrebbero sostituire dati del proprio sistema. In alcune sperimentazioni si sono usati semplici indici basati sulla media percentuale degli oggetti con almeno un errore in base ad opportune categorie di appartenenza (ad es., reticoli stradali, copertura del suolo,…).

Conclusioni e sviluppi futuri
Gli aspetti trattati sono determinanti per realizzare sia i singoli sistemi locali e sia quello globale della SDI e può essere utile in generale per normare il trsferimento dati anche in contesti non SDI. Rimane da affrontare il problema dell'impatto di avere soggetti che adottino risoluzioni diverse nei sistemi locali, la gestione degli oggetti globali che si spezzano al confine tra i gestori locali, le regole di armonizzazione al confine considerando anche il problema delle perturbazioni dovute alla computazione numerica.

Bibliografia

Belussi A., Liguori F., Marca J., Negri M., Pelagatti G. (2010), "Il Modello GeoUML. Regole di Interpretazione delle Specifiche di Contenuto per i Database Geotopografici", approvata dal comitato per le regole tecniche sui dati territoriali delle PA (www.DigitPA.it)
iso19107 (2003) ISO/TC 211, ISO 19107:2003, Geographic information - Spatial schema, text for FDIS, doc. N 1324
iso19125 (2004) ISO/TC 211. ISO 19125-1:2004 Geographic information - Simple feature access - Part 1: Common architecture, text for IS, doc. N 1563, 2004-01-23
Thompson R.J., Van Oosterom P.(2006), "Interchange of spatial data – inhibiting factors", 9th AGILE conference.