IMPORTANZA DEGLI STANDARD NELLA VALUTAZIONE DELLA QUALITÀ DEI DATI IN UN SISTEMA INFORMATIVO TERRITORIALE

Daniela CARRION(*), Federica MIGLIACCIO(**)

(*) DIIAR–Politecnico di Milano – P.zza L. da Vinci, 32 – 20133 Milano Tel. 0223996509 daniela.carrion@polimi.it
(**) DIIAR–Politecnico di Milano – P.zza L. da Vinci, 32 – 20133 Milano Tel. 0223996507 federica.migliaccio@polimi.it

Riassunto

Due punti fondamentali possono essere evidenziati, per quanto riguarda la valutazione, da parte di chi gestisca un Sistema Informativo Territoriale, dell'idoneità della base di dati relativamente alle applicazioni di interesse.

In primo luogo, la qualità di una base di dati georeferenziati può e deve essere valutata in base a una serie di parametri, contenuti in standard internazionali quali quelli definiti dai comitati tecnici ISO/TC 211: Geographic information - Geomatics e CEN/TC 287.

Un altro punto da tenere in considerazione è la necessità che i dati cartografici siano corredati da informazioni sufficienti perché l'utente possa giudicare se i dati gli siano davvero utili per i suoi scopi e se abbiano la qualità che li rende tali.

Il presente lavoro vuole avere carattere di semplice rassegna e riepilogo dei punti fondamentali relativi alla qualità dei dati, all'uso degli standard e agli organi internazionali che si occupano della loro definizione. Si evidenziano quali siano gli standard di interesse per la valutazione della qualità dei dati geografici, si sottolineano i vantaggi derivanti dal loro uso e dalla compilazione dei metadati. Su questi argomenti si veda anche Bianchin (2001).

Abstract

Two topics are to be taken into account to determine if the database of a Geographic Information System is suitable for the purpose of its users.

First of all the quality of a georeferenced database must be evaluated referring to parameters fixed by international standards defined by the technical committees ISO/TC 211: Geographic information - Geomatics and CEN/TC 287.

Another point is the need that cartographic data are accompanied by sufficient information to allow the users to judge if the data and the data quality fit their aims.

This paper is a simple review of the fundamental subjects related to data quality and the use of international standards. The importance of geographic data quality standards and of metadata will be highlighted. Another useful paper on this subject is Bianchin (2001).

Misura e valutazione della qualità dei dati

La preoccupazione relativa alla questione della qualità dei dati dipende da molti fattori, fra cui i principali sono i seguenti:

- l'uso crescente dei SIT nei processi di supporto alle decisioni implica che aumenti notevolmente la possibilità di soluzioni basate su dati di scarsa qualità;
- anche l'uso notevolmente aumentato di dati provenienti da fonti "secondarie" e non tradizionali (essenzialmente dovuto alla crescita di Internet) rende più probabile la diffusione di dati di bassa qualità;

- appare evidente come la crescente produzione e disponibilità di dati cartografici da parte del settore privato, nonché da parte delle agenzie nazionali, renda opportuna la conformità a standard internazionali per il controllo della qualità dei dati stessi.

Cerchiamo prima di tutto di rispondere ad alcune domande molto semplici ma fondamentali ai fini di stabilire le principali definizioni utili nell'ambito in esame.

Cos'è la qualità?

- La qualità può essere definita come il grado di eccellenza di un prodotto, di un servizio o di una prestazione.
- La qualità è un risultato altamente desiderabile, e può essere raggiunta attraverso una attenta gestione e controllo dei processi di produzione (controllo di qualità).
- Questi criteri possono essere applicati al concetto di qualità di una base di dati, dal momento che una base di dati è il risultato (in generale) di un processo di produzione, e l'affidabilità con cui tale processo è stato condotto ha un impatto sul valore e sull'utilità dei dati stessi.

Chi dovrebbe valutare la qualità dei dati?

Si possono in generale dare tre possibilità, corrispondenti ad altrettanti modelli di valutazione.

- Caso 1: modello "minimo"
  - il controllo e la valutazione di qualità vengono svolti a cura del produttore dei dati, basandosi su strategie di test di conformità per identificare dati che soddisfano soglie di qualità definite a priori;
  - è un approccio poco flessibile, il medesimo test in alcuni casi può rivelarsi troppo restrittivo, e in altri troppo lasco.
- Caso 2: uso di metadati
  - in questo modello l'errore è visto come un accadimento inevitabile, e non vengono imposti standard minimi a priori; è l'utente finale ad essere il responsabile della valutazione della "idoneità dei dati all'uso", mentre il produttore dei dati è responsabile della loro documentazione tramite i metadati;
  - questo approccio è flessibile, ma non è solitamente previsto che l'utente finale fornisca un riscontro ("*feedback*") al produttore di dati, quindi l'informazione viaggia "a senso unico" e il produttore non è in grado di correggere gli errori
- Caso 3: uso di standard e riscontro da parte dell'utente
  - il flusso delle informazioni è "bidirezionale", con un riscontro da parte dell'utente relativamente alle questioni della qualità dei dati; questo *feedback* è processato e analizzato per identificare i problemi significativi e gli interventi prioritari da avviare sui dati;
  - questo modello è utile non solo in un contesto di mercato, per assicurare che la base di dati corrisponda alle necessità e alle aspettative degli utenti e che il produttore possa intervenire per migliorarne la qualità.

Come misurare la qualità dei dati?

La qualità dei dati solitamente viene misurata in base a un insieme di indicatori (o parametri), di cui sotto si riportano qualli principalmente usati, con le loro definizioni e caratteristiche (Migliaccio, 2002).

- ***Accuratezza***

  E' l'inverso dell'errore, cioè della discrepanza fra il valore registrato nella base di dati e il valore "vero" (per quanto lo si può conoscere) o accettato come tale. La valutazione dell'accuratezza di una osservazione (misura) può essere fatta solo per confronto con la misura più accurata che sarebbe possibile ottenere. Il risultato è appunto la misura dell'errore:
  - per i dati di tipo numerico (essenzialmente, dati di posizione) vengono usati indici di tipo statistico, quali media e s.q.m.;

- per i dati di tipo non numerico (di solito, classificazioni) si possono valutare gli errori tramite la matrice dell'errore di classificazione.

Si possono inoltre definire:

- *Accuratezza spaziale*: accuratezza della componente spaziale di una base di dati; la metrica usata dipende dalle dimensioni delle entità in esame.
- *Accuratezza temporale*: concordanza fra il valore codificato e il valore reale delle coordinate temporali di un'entità; spesso per i dati di tipo geografico le coordinate temporali sono rappresentate dall'epoca alla quale l'entità era "valida", e in molti casi ciò vale contemporaneamente per tutta la base di dati.
- *Accuratezza tematica*: accuratezza dei valori degli attributi codificati in una base di dati; la metrica usata dipende dalla scala di misura dei dati.

- ***Risoluzione***

  Si riferisce alla quantità di dettaglio che è possibile distinguere nella componente spaziale, temporale o tematica; la risoluzione è sempre finita, perché non esiste un sistema di misura infinitamente preciso e perché comunque le basi di dati sono intenzionalmente generalizzate per ridurre il grado di dettaglio. In funzione dell'applicazione, non sempre un'elevata risoluzione è la scelta migliore; a volte una bassa risoluzione è più opportuna, se si vogliono ottenere modelli generali.

  Si possono definire:

  - *Risoluzione spaziale*: indica la più piccola differenza distinguibile fra due valori misurabili. Poiché la risoluzione spaziale è una informazione collegata al grado di dettaglio con cui i dati rappresentano la realtà, ci dice qual è il più piccolo oggetto distinguibile grazie ai dati a disposizione. Per i dati raster, la risoluzione è semplicemente definita dalla dimensione del pixel. Per i dati vettoriali, può essere definita come la minima dimensione dell'entità cartografica rappresentabile.
  - *Risoluzione temporale*: è la lunghezza (durata temporale) dell'intervallo di campionamento.
  - *Risoluzione tematica*: indica la precisione delle misure per un particolare tematismo. Per i dati quantitativi è analoga alla risoluzione spaziale; per dati qualitativi (categorie) rappresenta il dettaglio di definizione di una classe tematica.

- ***Scala***

  In assenza di altri dati sulla risoluzione spaziale e sull'accuratezza, la scala della cartografia di origine dei dati digitali può essere un parametro interessante. Infatti la scala di una carta contiene implicitamente informazioni sulla risoluzione spaziale (grado di dettaglio) dei dati. La linea più sottile che è possibile disegnare su una carta rappresenta un limite alla risoluzione raggiungibile a una determinata scala: nessun oggetto di dimensioni inferiori può essere registrato.

- ***Consistenza logica***

  Si riferisce al fatto che non ci siano dati in contraddizione fra di loro. Per valutare tale fatto si possono istituire

  - test di consistenza logica, per il controllo degli eventuali vincoli matematici o logici (relazioni matematiche o logiche fra i dati);
  - test di controllo dei vincoli topologici, per individuare ad esempio bordi mancanti o poligoni non etichettati.

  Anche la consistenza può essere definita in riferimento alle tre dimensioni (spaziale, temporale, tematica) dei dati geografici.

- ***Completezza***

  E' un parametro relativo alla mancanza di errori di omissione all'interno della base di dati, e si riferisce ai criteri usati per selezionare le informazioni da inserire fra i dati. La sua valutazione è basata su verifiche relative all'inclusione fra i dati di oggetti appartenenti a liste note oppure su verifiche relative all'inclusione fra i dati di oggetti di dimensioni (area o spessore) minime.

Ci sono due tipi di completezza:

- completezza dei dati, riferita agli errori di omissione; una base di dati è completa se contiene le informazioni relative a tutti gli oggetti definiti nelle specifiche;
- completezza del modello, riferita alla concordanza fra le specifiche della base di dati e il modello che è necessario definire per una particolare applicazione; una base di dati contiene un modello completo se le sue specifiche sono appropriate per una particolare applicazione.

Ovviamente, anche la completezza può fare riferimento alle tre dimensioni (spaziale, temporale, tematica) dei dati geografici.

Benefici portati dall'uso degli standard nella valutazione della qualità dei dati

In un SIT l'importanza dei dati è centrale: tipicamente, i dati in un SIT sono molto numerosi e tutti collegati gli uni agli altri.

Lo scopo di un SIT (come di qualsiasi sistema informativo) è ottenere risposte a specifiche domande: qualsiasi SIT deve essere in grado di produrre informazioni per dare risposte a queste domande e per condividere informazioni fra molti utenti. Peraltro, a differenza delle interrogazioni che si potrebbero avere in un ambito di lavoro di tipo industriale (dove la risposta è solitamente concentrata in pochi record), le interrogazioni fatte su una base di dati geografici coinvolgono moltissimi tipi di dati e di entità e inoltre danno spesso luogo anche alla produzione di cartografia.

Inoltre, la gestione dei dati di un SIT riguarda anche la condivisione dei dati stessi: l'interoperabilità consente di scambiare dati fra diverse organizzazioni o fra diverse applicazioni, con il risultato di generare e condividere informazioni più complete e più utili. Quindi, proprio perché le basi di dati dei SIT sono distribuite e dinamiche, gli standard e l'interoperabilità sono sempre stati un punto cruciale: il loro utilizzo dovrebbe aumentare sempre di più man mano che i SIT diventano non un semplice strumento per lavorare su singoli progetti, ma un ambiente di lavoro utile per scambiare informazioni fra diverse organizzazioni e all'interno della società.

La standardizzazione dei metadati è poi particolarmente importante perché consente agli utenti di comprendere l'informazione geografica scambiando "dati relativi ai dati". Gli standard dei metadati servono infatti per identificare e unificare le definizioni dei metadati che gli utenti dovranno utilizzare per potere condividere e riutilizzare dati geografici, aiutando a promuovere l'interoperabilità globale. Capire l'importanza dei metadati è fondamentale ai fini di costruire forti infrastrutture di dati spaziali.

Lo sviluppo di standard e regole tecniche da parte di organismi a cui sia riconosciuta sia da parte del settore pubblico che privato la necessaria autorità per svolgere queste attività è un elemento essenziale nell'infrastruttura tecnologica ed economica di una nazione, ed influenza grandemente la sua abilità competitiva e le strategie industriali.

La crescente globalizzazione ha infatti drasticamente cambiato il quadro internazionale e questo fatto, unito al ruolo in evoluzione della standardizzazione nel contesto europeo ed internazionale, rende necessario esaminare sia la forma che il contenuto delle procedure di standardizzazione, che comportano implicazioni non solo scientifiche e tecniche, ma anche economiche.

Organismi di standardizzazione a livello internazionale e standard per la qualità dei dati geografici

CEN/TC 287

Nel 1991, su proposta dell'AFNOR (Association Française de Normalisation), fu istituito uno specifico Comitato Tecnico nell'ambito del Comitato Europeo per la Standardizzazione (CEN, Comité Européen de Normalisation): CEN/TC 287 - Geographic Information.

Questo Comitato Tecnico finì i suoi lavori nel 1999, risultanti in un insieme di "Standard europei sperimentali" (ENV) nel campo dell'informazione Geografica (si veda la Tabella 1), di cui quelli relativi alla qualità dei dati e ai metadati sono il 12656 e il 12657. Tali standard sono stati peraltro recentemente ritirati e sostituiti da corrispondenti standard EN ISO.

Numero CEN	Nome	Anno
ENV 12009	Geographic Information - Reference Model	1997
ENV 12160	Geographic Information - Data description - Spatial schema	1997
ENV 12656	Geographic Information - Data description - Quality	1998
ENV 12657	Geographic Information - Data description - Metadata	1998
ENV 12658	Geographic Information - Data description - Transfer	1998
ENV 12661	Geographic Information - Referencing - Geographic identifiers	1998
ENV 12762	Geographic Information - Referencing - Position	1998
ENV 13376	Geographic Information - Data description - Rules for application schema	1999

Tabella 1 Standard ENV pubblicati dal CEN

ISO/TC 211

L'Organizzazione Internazionale per la Standardizzazione (International Organisation for Standardisation) fu fondata nella primavera del 1995. ISO non è però un acronimo, ma deriva dalla parola greca che significa "uguale", per indicare la cooperazione di diversi partner che utilizzano standard nello scambio di dati. Quando l'organizzazione ISO fu fondata, ISO/TC 211 e CEN/TC 287 si accordarono perché CEN/TC 287 finisse il suo programma senza ulteriori sviluppi e contribuisse nel definire temi di lavoro per il Comitato ISO/TC 211.

ISO/TC 211 ha un programma che copre una vastissima tipologia di standard, pari a ben 40: molti di questi sono ormai completati e pubblicati. Si veda la Tabella 2 per quelli relativi alla qualità dei dati geografici.

Molti sostennero che a quel punto il Comitato ISO/TC 211 avesse assorbito il programma europeo per la standardizzazione, e che non ci fosse più bisogno di avere due comitati tecnici. Questo fu vero quando iniziarono i lavori di ISO/TC 211, ma ora che gli standard internazionali sono stati pubblicati, l'Europa ha anche bisogno di decidere come adottarli e applicarli in ambito europeo (Carosio et al., 2001).

Numero ISO	Nome	Anno	EN ISO
ISO 19113	Geographic information - Quality principles	2002	EN ISO 19113:2005
ISO 19114	Geographic information - Quality evaluation procedures	2003	EN ISO 19114:2005
ISO 19115	Geographic information – Metadati	2003	EN ISO 19115:2005

Tabella 2 Progetti ISO/TC 211 relativi alla qualità dei dati geografici

Seguendo la proposta del Technical Board del CEN, numerata BSI IST/36, le attività del CEN/TC 287 devono essere riattivate sotto il segretariato del NEN (Istituto di Normalizzazione dei Paesi Bassi) a partire dal 2003, con lo scopo di portare all'armonizzazione degli standard CEN/TC 287, sviluppati fra il 1992 e il 1999, e l'insieme di standard ISO/TC 211 (Geographic Information/Geomatics), sviluppati a partire dal 1995 (spesso indicati come ISO 191xxx).

Subito dopo la ricostituzione di CEN/TC 287, il Comitato Tecnico decise di riformulare i suoi compiti (Risoluzione 40 del meeting di Delft del Novembre 2003, che riportiamo secondo la sua formulazione originale):

"Standardisation in the field of digital geographic information for Europe:

The committee will produce a structured framework of standards and guidelines, which specify a methodology to define, describe and transfer geographic data and services. This work will be carried out in close co-operation with ISO/TC 211 in order to avoid duplication of work.

*The standards will support the consistent use of geographic information throughout Europe in a manner that is compatible with international usage. They will support a spatial data infrastructure at all levels in Europe"***.**
Il processo di revisione degli standard ENV esistenti dovrebbe indicare se sia opportuno applicare direttamente gli standard ISO 19xxx o se invece non siano necessari specifici profili "europei", influenzando così la direzione da prendere relativamente allo sviluppo degli standard in Europa.

Open GIS Consortium
Il Consorzio Open GIS fu fondato per iniziativa di aziende, enti ed università nello stesso periodo in cui vedeva la luce il Comitato ISO/TC 211. Poiché si poneva gli stessi obiettivi di tale Comitato e presentava una notevole sovrapposizione dell'ambito di lavoro, nel 1998 fu steso un accordo di cooperazione. In seguito ad esso, l'OGC ha adottato gli standard ISO come specifiche astratte e ha sviluppato specifiche di implementazione che vengono presentate all'ISO e possono poi diventare standard internazionali. Ciò significa comunque che l'OGC non sviluppa propri standard di qualità.

Conclusioni
- La qualità dei dati definisce il grado di eccellenza di una base di dati e viene valutata relativamente alle specifiche della base di dati, che definiscono il livello voluto di generalizzazione e astrazione. Volendo, si possono valutare anche la qualità di queste specifiche e la loro adeguatezza per determinate applicazioni.
- La valutazione della qualità dei dati è svolta in funzione di diversi parametri, inclusi l'accuratezza, la risoluzione, la consistenza e la completezza. Ogni componente di una base di dati geografica può essere valutata secondo le dimensioni di spazio, tempo e tematismo (essendo queste le tre dimensioni fondamentali dei dati geografici).
- Nell'ambito della gestione dei dati di un SIT in maniera condivisa e per facilitare lo scambio di dati fra diverse applicazioni (interoperabilità) gli standard svolgono un ruolo cruciale: il loro utilizzo dovrebbe aumentare quanto più i SIT diventano un ambiente di lavoro utile per scambiare informazioni.
- Esiste un vasto lavoro da parte di organismi internazionali (ISO e CEN) preposti alla standardizzazione delle informazioni geografiche, dei parametri per la valutazione della loro qualità e dei metadati, che è importante conoscere.

Bibliografia
Bianchin A. (2001), "Nuovi approcci alla validazione dei DB cartografici", *Atti della 5^a conferenza nazionale ASITA*, Rimini, 9-12 Ottobre 2001, Vol. I, V-XVI
Carosio A., Nocera R. (2001), "Norme tecniche internazionali per l'informazione geografica. Il ruolo dell'Europa", *Atti della 5^a conferenza nazionale ASITA*, Rimini, 9-12 Ottobre 2001, Vol. I, 393-398
Migliaccio F. (2002), "Qualità dei dati e metadati all'interno di un GIS", *Atti della 6^a conferenza nazionale ASITA*, Perugia, 5-8 Novembre 2002, Vol. II, 1551-1556

http://www.ertico.com/links/gdf/gdfdoc/gdfdoc.htm
http://www.isotc211.org
http://www.ncgia.ucsb.edu/giscc/units/u100/u100_f.html
http://www.uni.com/it/

Ringraziamenti
La presente nota è stata redatta nell'ambito dei lavori relativi alla ricerca COFIN 2004 Strutture evolute della cartografia numerica per i GIS e l'ambiente WEB, Coordinatore Scientifico: prof. Riccardo Galetto, Unità Operativa del Politecnico di Milano.