How to Improve Geospatial Data Usability: From Metadata to Quality-Aware GIS Community

R. Devillers[1,2], Y. Bédard[2,3], M. Gervais[2], R. Jeansoulin[4,2], F. Pinet[5], M. Schneider[6], L. Bejaoui[2,3], M.-A. Levesque[2,3], M. Salehi[2,3], A. Zargar[1]

1 – Department of Geography, Memorial University of Newfoundland, St. John's (NL), Canada
2 – Centre for Research in Geomatics (CRG), Laval University, Québec (QC), Canada
3 – Industrial Research Chair in Geospatial Databases for Decision Support, Laval University, Québec (QC), Canada
4 – Laboratoire d'informatique, Institut Gaspard-Monge, U. de Marne-la-Vallée, France
5 – Cemagref-Clermont-Ferrand, France
6 – Department of Computer Sciences, Université Blaise-Pascal, Clermont-Ferrand, France

INTRODUCTION

The field of Geomatics/GISciences has witnessed major changes since its origins in the 1960s. If developments in this field first looked at ways to transfer and store spatial and semantic information from paper documents into computers (e.g. spatial data structures - raster vs vector; scan; topology), the focus moved later to the design of more advanced ways to store, retrieve and analyze geospatial data. The field grew exponentially and, in the last two decades, organisations started to realise that large volumes of geospatial data were produced, but mostly remained unknown from many potential users (even within a same organisation). In order to have a better Return on Investment (ROI), and to encourage an increased use, organisations and countries started to develop initiatives like Digital Libraries and Spatial Data Infrastructures. This moved a lot of the research focus from the systems themselves to data transfers and reuse issues (e.g. interoperability, metadata, ontologies, data fusion). We are nowadays entering the next phase: the widespread usage of these data by people who haven't collected or integrated them and may not have been initially targeted as primary users. With the increasing ease of access to geospatial data and with the user-friendliness of today's Geographic Information Systems (GIS) and related web-based solutions, geospatial information is more than ever reaching the hands of the general public (e.g. Google Earth, Virtual Earth). Similarly, expert users in different fields of applications have also seen their number increasing by orders of magnitudes worldwide. Digital Libraries and Spatial Data Infrastructures are now facing huge downloads. For instance, the Canadian Web portal *Geobase* that provides free access to geospatial data (e.g. DEM, road network), had an increase from about 210,000 downloads in 2003-2004, to about 600,000 in 2004-2005, and more than 2,200,000 downloads in 2005-2006 (Geomatics Canada, personal communication, 2007). Since April 2007, the Canadian Government increased the amount of data available for free by providing for free their entire National Topographic Database (NTDB). In addition to this increasing access to geospatial data, GIS applications are not anymore restricted to traditional land/resources-related uses but now reach most disciplines, ranging from Science/Engineering to Human/Social or Medical Sciences. Consequently, most of the new users have limited or no knowledge of the geospatial field and the underlying nature of spatial referencing (e.g. reference systems/projections, scale, generalisation, accuracy). Furthermore, it appears that problems faced by users from the general public are also emerging more often than ever before. Similarly, users having an expertise in geomatics often just cannot know the key characteristics which are necessary to assess the usefulness of the data being downloaded, nor the added uncertainty resulting from the integration of such datasets. Consequently, an increasingly important research agenda is now to make sure the level of usability of spatial data is better known for contexts that were not always planned when the data were collected. One objective when willing to improve spatial data

usability is to try to reduce the risks of misuse of these data and the risks of potential accidents that could result from these misuses.

This research agenda is not completely new as research has been going on for years on ways to document datasets using metadata (e.g. FGDC and ISO 19115 standards), assess and visualize their uncertainty, etc. But such researches are definitively more active and relevant than ever in this new context of geospatial data usage. Furthermore, if several academic works were related to this topic, very little practical advances (e.g. in commercial software, professional practices, system design methods or legal documents) were done to really support end-users in the assessment of data usability and in dealing with the potential consequences of misuses.

Data usability can be generally defined as the capacity of data to be used by a given user (individuals, groups, organisation), for a given purpose, a given area, and a given epoch. This encompasses very different issues, such as being able to know that some data exist, being able to access them (e.g. access, interoperability, cost, privacy), understanding them (e.g. ontologies, design), understanding their limitations, making use of the data, etc. A first group of fifty-two elements related to spatial data usability were identified after a first workshop on this issue in 2001. This list, that is probably not exhaustive, includes elements typically found in quality standards like ISO 19113 or 19115, that are identified as the internal data quality elements (e.g. positional accuracy, completeness, logical consistency). But these usability elements also include issues related to the wider external data quality (e.g. accessibility, availability, cost, reliability), typically identified as defining the external data quality (Devillers and Jeansoulin, 2006).

This paper looks at different ways that were used in the last decades to improve spatial data usability. We first present a view of this evolution. We then discuss different researches done by the authors during the past 10 years in North-America and Europe to provide preliminary solutions to the problem of assessing spatial data usability and reducing data misuse. We place these works into a larger context, see how they are complementary, their limitations, and then discuss future research directions that should help the end-users in the highly complex task of assessing the usability of their data.

FROM METADATA TO SPATIAL DATA QUALITY VISUALISATION

A main challenge, when willing to improve spatial data usability, relies on being able to transfer some knowledge, or expertise, of the data producers to the end-users of these data. This is done with the assumption that such knowledge should ultimately allow end-users to better understand the characteristics and limitations of the data, and then assess the general usability of the data in their context.

Historically, this transfer of knowledge has been done, and still is, using metadata (i.e. data about data). We should note that it is still very typical to see datasets with no metadata at all. Metadata are produced by data producers to qualify or quantify certain aspects of the data (e.g. describing the spatial and temporal coverages, the reference system, scale, spatial accuracy, completeness). If metadata were initially only a digital version of the information recorded in the cartouche of paper maps (e.g. source, scale, projection), they evolved to include other aspects related to the new digital medium (e.g. data format, access). Having metadata is obviously better than having no information, but metadata proved to be of very limited help for the end-users. In addition to being either missing, incomplete, or too general to be of any real help (e.g. describing the average accuracy of all objects located on all maps from a data collection), metadata are more a technical documentation of data characteristics made by the producers for internal management needs, than a product really designed to help the end-users (e.g. Timpf *et al.*, 1996; Devillers *et al.*, 2007). The multidimensional nature of metadata also increases their complexity as end-users have to understand many metadata that can be inter-related. As a consequence, metadata are usually too complex to be understood by the end-users

and are then most of the times neglected by them. If recent metadata standardization initiatives such as the ISO 19115 (ISO/TC 211, 2003) are likely to increase the documentation of datasets and facilitate the communications between systems, it may however have worsened the problem by making the metadata more technical than ever, often replacing information typically described in free text by text with a very specific format. Producing good metadata can be very challenging for a data producer. Metadata can easily end-up to take more space than the data themselves in a system and can then be created at a higher cost than data production. As a consequence, these standards are typically used by data producers to document a minimal set of metadata, identified as "discovery metadata". Such metadata become very useful when searching for datasets in a digital library, but is too limited to really support end-users in their assessment of the usefulness of a dataset for their specific need.

Realising these limitations, different initiatives developed ways to visualise some of these metadata in a more traditional cartographic way (e.g. using thematic map that can show the different qualities using different colours or symbols) (see for instance Devillers and Beard, 2006 for a discussion of these works). Such an approach is interesting but usually provides a limited view on the overall data quality, as these works usually focus on only one type of quality parameter to visualise (e.g. positional accuracy). In addition, they usually miss the complexity of metadata as they usually do not address the problem of the different levels of details of metadata (e.g. metadata describing an object instance, like a specific building *vs.* an object class, like all the buildings *vs.* a whole dataset, like the building, roads, rivers, etc.).

FROM VISUALISATION TO QUALITY-AWARE GIS COMMUNITY

Transferring some knowledge from the data producers to the data users is somehow a typical communication problem that can find its theoretical foundation in the general communication theory. Part of the efficiency of any communication between two persons relies on how close their frames of reference are. Frames of reference involve for instance the language they use, their knowledge, and their life's experience in general. Hence, two persons that would not share a common language would for instance have distinct frames of reference that would not allow an efficient communication. To simplify, the different solutions proposed over time to communicate quality information aimed at bringing data producers and users' frames of reference closer to help them understanding each other. This has been done in three different ways (cf. figure 1) through:

A. Asking data producers to formalize their implicit knowledge of geospatial data characteristics and communicate it in a more understandable way for a given set of users (cf. A on Figure 1):
   - This first aspect has mainly been addressed, with very limited success, by metadata. In addition to the difficulty of communicating such technical information, data producers are limited in their budgets and usually cannot afford having metadata costing more than data collection itself. Data producers than limit the level of detail provided in their metadata and then often limit their usefulness.

B. Asking data users to expand their knowledge of geospatial data to better understand data characteristics and limitations (cf. B on Figure 1):
   - This second aspect would be hard to achieve as it means end-users need to become experts in geomatics, or to be at least aware of many issues, to be able to use any geospatial data with limited risk. But some basic advises provided to users before they use a GIS could probably help decreasing the risks of misuse (e.g. mentioning the presence of uncertainty related to objects positions' on a map, or the possibility that some objects may be missing from the map).
   - Attempts were also made to help users aggregating fine-grained metadata in order to assess quality indicators at the attribute and geometry levels, then up to the object class level and ultimately to the dataset level (e.g. Bédard & Vallière 1995), to finally

compare the results with their needs. However, such an approach proved to be too complex.

C. Asking a third party to act as a mediator between data producers and data users (cf. C on the Figure 1) – this could be compared to a language translator in a traditional human communication process:

- This third aspect suggests making the link between data producers and data users using a third party that could improve the communication efficiency. This third party could be human (e.g. an expert) or software. As the experience showed that data producers are not sufficiently efficient in their communication, and as typical end-users cannot understand sufficiently technical details about the data to limit their risks of misuse, this solution appeared to be offering potential solutions to this problem. Hence, most works presented in this paper focused on this third solution through the use of different types of human or software mediators that would try to fill the missing link between data producers and users.
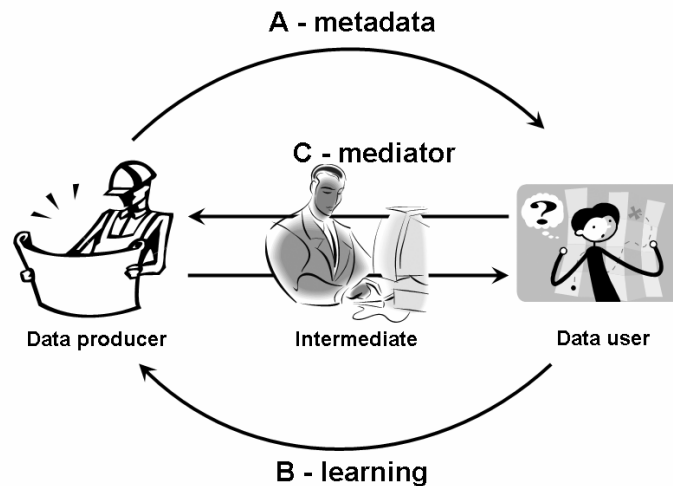

Figure 1: Different approaches that can be used to improve data users understanding of geospatial data usability.

The different research solutions tested over the years to improve the usability of spatial data are represented on the figure 2, showing where they fit into the whole flow of spatial data, from their production to their use in a context of decision-making. On this figure, the approaches are divided into two classes: those improving the data or the systems before the data are distributed (i.e. *a priori* approaches), and those trying to improve the selection or the use of the data after their production (i.e. *a posteriori* approaches).
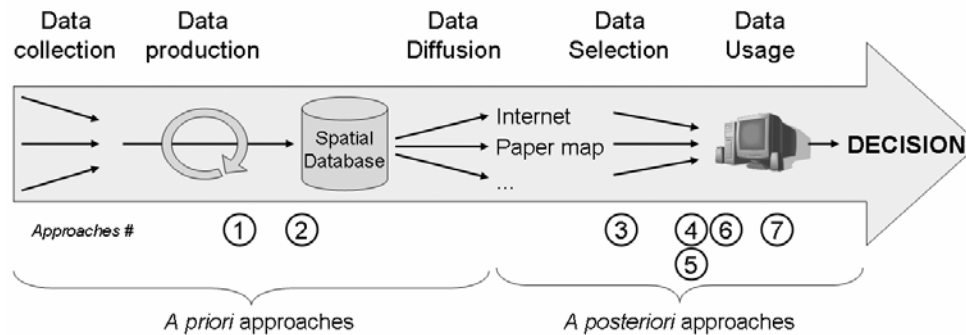
Figure 2: Location of the different approaches in the general flow of spatial data.

A priori approaches:

1. The first approach looks at improving spatial integrity constraints. Usability relies, amongst other things, on the confidence one may have of the internal quality of spatial data. Metadata provide a documentation of the data, but producers and end-users must be sure that the data comply with this documentation. Spatial integrity constraints aim at evaluating and maintaining data quality. They ensure the data comply with a set of rules specified by the data producer. Some of the authors first developed a prototype to describe spatial integrity constraints using a combination of semi-structured natural language and the ISO e-relate 3x3 matrix for spatial databases (Bédard *et al.*, 1998; Normand, 1999). They are now:

    *   extending this approach for spatial datacubes, as those used in SOLAP, Spatial Data Mining and Spatial Dashboards (Salehi *et al.*, 2007); Spatial datacubes are aimed at supporting geographic knowledge discovery as well as certain types of spatial decision-making. In traditional spatial databases, spatial integrity constraints have been employed to improve internal quality of spatial data. However, spatial datacubes require additional integrity constraints in comparison to the traditional databases found in transactional GIS systems. These extra constraints concern the supplementary information included in these datacubes, such as spatial dimensions and hierarchies, aggregated data, multidimensional cross-tabulation of data, and the existence of a temporal dimension with several levels of granularity. This project deals with the classification of these integrity constraints and building proprietary integrity constraint specification languages tailored for geospatial datacubes. The result will improve internal quality of spatial datacubes, and hence, the final quality of the decision making process.
    *   extending this approach for fuzzy objects and are designing fuzzy spatial integrity constraints to obtain richer quality information (Bejaoui *et al.*, 2007);
    *   extending the Object Constraint Language (OCL) for spatial integrity constraints (on-going PhD Thesis, M. Duboisset; Duboisset *et al.*, 2005, Pinet *et al.*, 2007). At present, the standard constraint language is the Object Constraint Language (OCL), recognised by the ISO/TC 211. It is an important part of the Unified Modeling Language (UML), accepted both by the industrial domain and the scientific community. This work consists in integrating spatial features into OCL to model spatial integrity constraints. A final goal of this work is to allow designers to specify spatial constraints in OCL, independently of the platforms, and then to generate equivalent integrity checking mechanisms into different relational DBMS;

2. Designing a way to identify risky uses of geospatial data and embed this knowledge into the mapping software (Spatial OLAP in this case) during the design of the database, to warn end-users of potential risks when manipulating geospatial data (Levesque *et al.*, 2007b). This approach can be classified as *a priori* because the professional's role is to identify risks during the database development process, to document those risks and to decide jointly with the user best management strategy to adopt against those risks (e.g. trying to reduce the risk *vs* avoiding it *vs* absorbing the remaining risk and proceed with the decision to be made). For each risk identified, professional and user have to fill a form in which the risk is described and the management strategy is chosen. Users must sign each form to confirm that they know the risk and that the strategy adopted is good for them and finally, that they are comfortable with the residual risk. The risk management model developed by Levesque (2007b) is inspired from the ISO standards for risk management and security information;

A posteriori approaches:

3. Designing a progressive "impedance analyser", to help transformation of data queries (selection, re-engineering), in order to better meet users' needs (Guemeida, 2007). This follows the present semantic web trend that promotes a widespread use of mediation between sources and users, trigered by the user queries. Series of software agents are translating partial views from the user's requirements (global ontology), onto the specific (local) schemata of either GI catalogues, metadata, or map servers in the OGC Web services architecture (Lassoued, 2005);

4. Combining visualisation and decision-support techniques to provide end-users with a system that could support them in their understanding of spatial data quality (Devillers *et al.*, 2007; Levesque, 2007a). This expands the works done on spatial data quality visualisation by offering a more complete spatial data quality information system. The works from Devillers et al. (2007) for instance integrate existing metadata into a multidimensional database (such as the ones used by OLAP systems). Such a structure allows for instance to organise metadata according to the level they describe (i.e. dataset, feature class, feature instance), as well as preserving the multidimensional nature of spatial data quality (both their diversity and hierarchy). Users can then browse into these metadata at different levels of detail within a typical GIS interface in order to better understand the potential problems that can be related to the use of a given dataset. Such an approach relies however on the available metadata and would then be limited in many cases in practice as many datasets have little or no metadata;

5. Collecting and formalising the knowledge experts have of geospatial datasets and getting their opinions on the ability of datasets to be used with limited risks for given applications (Levesque, 2007a). This approach is complementary to the previous one as it is another way to collect additional metadata. It uses a top-down approach (from the expert knowledge down to the metadata, whereas the previous approach used a bottom-up approach, going from metadata up to their visualisation). Such an approach can however be hard to apply in some cases as it involves collecting data from experts, formalising them, integrating conflicting information, dealing with privacy issues, etc;

6. Recommending end-users to request the opinion of an expert in spatial data quality (a *quality auditor*) that could proceed to the complex task of evaluating how different datasets can be used in a specific context (Gervais, 2004). Such approach would comply notably with legal requirements in the countries studied, follow strict methods adhered by registered professionals and involve professional liabilities and insurances, that is an approach based on a true professional act. To justify their opinion, the professionals would need to get enough information on the data quality. After analysing the data, quality auditors could provide a *Quality Certificates* that would present their conclusions (Gervais - in progress). This approach has some good

advantages for the user. For example, legal rules in civil law and common law impose to the professional to act in conformity with high ethic standards. Secondly, this approach has the advantage to transfer the final liability from the data producers to the professional. To accomplish their duties, the professionals will have to explore and understand the users' needs and the use context. Based on this information, the professional will have to provide recommendations, relevant advices and specific warnings related to the use expected, all into the quality certificate. The quality certificate is similar to many reports produced by professionals in standard business. The extent of this type of report will depend of the context and the user's needs.

7. Designing a system that could extend existing GIS to link GIS tools to spatial data quality information, in order to have the GIS providing warnings, error estimates, etc. to the end-users when manipulating uncertain data (A. Zargar, MSc thesis in progress). This project involves understanding how certain types of uncertainty can affect certain types of GIS functions. By knowing this, and for a limited number of functions, it would be possible to use the knowledge of the spatial data quality stored in the metadata and to link it to the functions, in order to provide warnings or uncertainty estimates to the end-users that could help him during the interpretation of the data;

All these approaches would help improve spatial data usability, but all have advantages and drawbacks. Combining these different approaches would however improve significantly our capacity to understand the characteristics and limitations of the datasets available for use.

CONCLUSIONS AND FUTURE DIRECTIONS

The Geomatics/GISciences discipline has evolved rapidly in the past forty years and one of the main challenges is now to help geospatial data end-users to assess the usability of datasets collected from third parties. This is a highly complex task and we don't think any perfect solution will ever exist. However, the current situation could be easily improved. Different complementary approaches have been explored and will be explored in the future to address this problem. These works can be grouped into approaches that try to (1) improve tools that can validate the quality of datasets, (2) collect information about datasets that could complement metadata, (3) transform metadata into an information easier to understand by end-users, (4) enhance existing mapping software to warn users of potential risks, (5) improve querying techniques to help users access the most appropriate data, and (6) findings professional and legal processes to absorb the remaining uncertainty. All of these approaches have advantages and drawbacks but are complementary. Trying to assess geospatial data usability is a highly complex task as it involves comparing, on one hand, data characteristics that can be missing, incomplete, inaccurate or out-of-date, with, on the other hand, users requirements which users themselves have a lot of difficulty to assess. Future works are likely to move their focus from uncertainty assessment and reduction techniques to uncertainty absorption ones (i.e. finding ways to work with a given remaining uncertainty). But it is likely that the increasing access to geospatial data combined with the increasing use by non-expert users will increase the importance of this problem in the next decade.

BIBLIOGRAPHY

Bédard, Y. and Vallière, D., 1995 Qualité des données à référence spatiale dans un contexte gouvernemental, Rapport de recherche pour le Plan géomatique du gouvernement du Québec (PGGQ), 55p.

Bédard, Y., Normand, P. and Larrivée, S., 1998 Modélisation des contraintes d'intégrité spatiale, Research Report for the Mapping Service, Quebec Ministry of Natural Resources, 76 p.

Bejaoui, L., Bédard, Y., Pinet, F., Schneider, M., 2007 Logical consistency of vague spatio-temporal objects and relations, Proceedings of the 5th International Symposium on Spatial Data Quality, Enschede, The Netherlands, June 2007.

Devillers, R., and Beard, K., Communication and Use of Spatial Data Quality Information in GIS. In R. Devillers and R. Jeansoulin (eds.). Fundamentals of Spatial Data Quality. ISTE Publishing: p.237-253, 2006.

Devillers, R., Bédard, Y., Jeansoulin, R. and Moulin, B., Towards Spatial Data Quality Information Analysis Tools for Experts Assessing the Fitness for Use of Spatial Data. International Journal of Geographical Information Sciences, 21(3): p.261-282, 2007.

Devillers, R., and Jeansoulin, R. (eds.), 2006 Fundamentals of Spatial Data Quality, ISTE Publishing, London, UK.

Duboisset, M., Pinet, F., Kang, M.A. and Schneider, M. 2005 Precise modeling and verification of topological integrity constraints in spatial databases: from an expressive power study to code generation principles. Lecture Notes in Computer Science vol.3716, Springer : pp.465-482.

Gervais, M., 2004 La pertinence d'un manuel d'instructions au sein d'une stratégie de gestion du risque juridique découlant de la fourniture de données géographiques numériques. PhD thesis, Université Laval (Canada) and Université de Marne-la-Vallée (France), pp 344.

Guemeida, M., Jeansoulin, R. and Salzano, G., 2007 Quality-Aware Metadata-based Mediator for Environmental Information. International Symposium on Spatial Data Quality, Enschede, The Netherlands, June 2007.

ISO/TC 211, Geographic Information - Metadata 19115, 2003.

Lassoued, Y., 2005 Médiation de la qualité dans les systèmes d'information géographiques. PhD thesis, Université de Provence, Marseille (France).

Levesque, J., 2007a Évaluation de la qualité des données géospatiales – Approche top-down et gestion de la métaqualité. MSc thesis, Université Laval (Canada), 125 p.

Levesque M.-A., Bédard, Y., Gervais, M. and Devillers, R., 2007b Towards a safer use of spatial datacubes: communicating warnings to users. International Symposium on Spatial Data Quality, Enschede, The Netherlands, June 2007.

Normand, P., 1999 Modélisation des contraintes d'intégrité spatiale: théorie et exemples d'application. MSc thesis, Université Laval (Canada), 95 p.

Pinet, F., Duboisset, M. and Soulignac, V., Using UML and OCL to Maintain the Consistency of Spatial Data in Environmental Information Systems. Environmental Modelling & Software, 22, 2007.

Salehi, M., Bédard, Y., Mostafavi, M.A., Brodeur, J., 2007 From Transactional Spatial Databases Integrity Constraints to Spatial Data Cubes Integrity Constraints, Proceedings of the 5th International Symposium on Spatial Data Quality, Enschede, The Netherlands, June 2007.

Timpf, S., Raubal, M. and Kuhn, W., Experiences with Metadata, Proceedings of Symposium on Spatial Data Handling, SDH'06, Advances in GIS Research II, Delft, The Netherlands, August 12-16th, IGU, p.12B.31-12B.43, 1996.