Towards a Quantitative Evaluation of Geospatial Metadata Quality in the Context of Semantic Interoperability

Tarek Sboui, Mehrdad Salehi & Yvan Bédard

Abstract

Semantic interoperability is a process to facilitate the reuse of geospatial data in a distributed and heterogeneous environment. In this process, the provided geospatial metadata that are appropriate for the intended use may be incomplete or not appropriate for data reuse. Thus, the external quality (fitness for use) of these metadata seems important for data reuse, since it has the potential to protect re-users from the risk of misinterpreting geospatial data. In this paper, we aim to provide a step forward in making re-users aware of the quality of geospatial metadata. We introduce a set of indicators for geospatial metadata quality and we propose a method to evaluate them with respect to the context of re-users. Based on this evaluation, we derive warnings that indicate the degree of risk of data misinterpretation related to metadata quality in the context of semantic interoperability.

1 Introduction

Over the last decades, there has been an exponential increase in the amount of geospatial data available from multiple sources. Reusing this data can significantly decrease the cost of geospatial application (Fonseca *et al.*, 2000). In order to develop ways to enhance the reuse of available geospatial data, significant research efforts have been carried out. Among these efforts, semantic interoperability has been extensively investigated (e.g., Bishr, 1998; Harvey *et al*., 1999), but it still remains a challenge in spite of all these efforts (Staub *et al.*, 2008). For facilitating geospatial data use/reuse, providers and application designers tag data with additional information called *geospatial metadata*. Metadata can be thematic (e.g., data acquisition method), spatial (e.g., spatial reference system used) or temporal (e.g., the time of data acquisition). The purpose of geospatial metadata is to facilitate the interpretation of data. For example, a map including a set of geometric representations (data) can be accompanied by a legend (geospatial metadata) that facilitates the interpretation of the map. However, geospatial metadata which are appropriate for a specific application may be less appropriate for another: for example, the dates of certain photographs displayed on Google Earth are often one year old, which may have no impact on several usages but may mislead others. In fact, the quality of geospatial metadata may be insufficient for certain data re-usage. Poor geospatial metadata quality may cause a risk of misinterpreting data, and may undermine the reuse of geospatial data (Agumya and Hunter, 2002), i.e., the main aim of semantic interoperability.

In order to effectively respond to the risk of misinterpreting geospatial data, re-users need to be aware of geospatial metadata quality with regard to their application. The aim of this paper is to provide a step forward in making re-users aware of the quality of metadata. We propose a method to evaluate a set of indicators for geospatial metadata quality. Such evaluation will help users to appropriately reuse geospatial data considering semantic interoperability. For example, if the quality of geospatial metadata is very low, the risk of data misinterpretation will be higher and, consequently, it may not be advisable to reuse the geospatial data.

The contributions of this paper are as follows: 1) we discuss the risk of misinterpreting data due to the lack of information about the quality of geospatial metadata in the context of semantic interoperability; 2) we introduce a set of indicators that have a major role in indicating the quality of geospatial metadata; 3) we propose a method to evaluate the quality of metadata that includes an evaluation function for each indicator and manners to calculate these functions; and 4) we use the result of quality evaluation with a framework and a set of warnings to inform re-users about the risk of data misinterpretation related to metadata quality.

In Section 2 we discuss the risk of data misinterpretation related to the semantic interoperability of geospatial data. Following this, we propose an evaluation of the quality of metadata to respond to the risk of data misinterpretation using appropriate warnings. We conclude and present further works in Section 4.

2 Semantic interoperability of geospatial data and the risk of misinterpretation

2.1 Semantic interoperability of geospatial data

Semantic interoperability has been defined as the ability of different systems to exchange data and applications in an accurate and effective manner (Bishr, 1998; Harvey *et al.*, 1999; Brodeur, 2004). The principal aim of semantic interoperability is to reuse data and applications located in different sources.

Semantic interoperability has been viewed as the technical analogy to the human communication process (Brodeur, 2004; Kuhn, 2005). In this process, agents (a source/provider and a receiver/re-user) communicate messages (i.e., data) which are a set of organized terms. The communication process works properly when the re-user interprets data with the meaning that was originally intended by the provider (Shannon, 1948; Schramm, 1971; Bédard, 1986). To do so, re-users typically need metadata that describes the content of data regarding a specific application. For reason of simplicity, we consider that metadata consist of a set of elements. An element can be a word/phrase, a set of graphic symbols, or a combination of both. Annotation mechanisms for metadata are another key aspect but are beyond the scope of this paper. Figure 1 shows a snapshot of a communication process that involves only two interveners (e.g., source data producers and user's destination). Perfect communication takes place when A1 and A2 correspond exactly to each other. We should notice that such a process could also involve a chain of multiple interveners (e.g., mediators, translators, aggregators), each with their own interpretation.

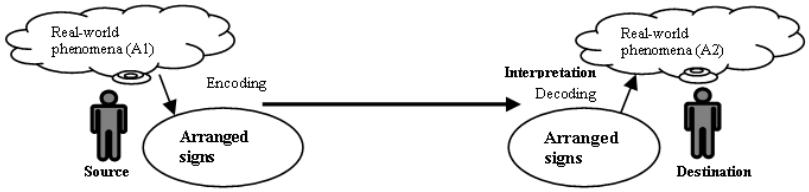

Figure 1. Sign interpretation within the communication process.

2.2 Risk of geospatial data misinterpretation related to semantic interoperability

In semantic interoperability, data may be reused in an application that is different from the one originally intended. In this process, metadata that were considered appropriate for the original application with regard to completeness, trust, clarity and levels of detail may be considered less appropriate for the second one. The appropriateness of metadata for a given application can be referred to as the external quality of metadata (fitness-for-use). Although such quality is important, evidence shows that users typically do not have a proper understanding of it (Agumya and Hunter, 2002). Consequently, they likely make wrong assumptions about the external quality of metadata. Such assumptions have the potential to expose re-users to the risk of misinterpretation. This risk is characterised by the probability of making faulty interpretations, and by the damage that can be caused by such misinterpretation.

We illustrate the risk of data misinterpretation with the following example: an agent provider sends a message, i.e., data tagged with metadata, to an agent re-user. Data consist of a set of line segments free of any intrinsic signification. We suppose the data provider is aware of eventual road-network analysis applications where data can be used and consequently tagged metadata consisting of these three elements: (1) lines represent roads, (2) bold lines indicate segments with restrictions for vehicles, and (3) road intersections have no turn restrictions unless indicated. This metadata is originally intended for a civil road network application, and is reused, by means of semantic interoperability, for an emergency road network application. Without more complete, clear or detailed indications, the semantic quality of geospatial metadata may be insufficient and lead re-users to false assumptions. This includes, for this example, that bold lines represent restrictions for all vehicles, including emergency vehicles, all line intersections are road intersections, while there could be viaducts and consequently no possibility to turn in spite of the absence of explicit restrictions in the dataset, all roads are included in the dataset while only public roads are, and so on. Such assumptions may cause a risk of faulty interpretations and usages of data. In order to effectively respond to these risks, re-users need information about the quality of metadata regarding their application.

3 Evaluating the quality of geospatial metadata in semantic interoperability

Significant research efforts have been carried out to evaluate and enhance the quality of geospatial data (e.g., Agumya and Hunter 2002; Frank *et al.*, 2004; Devillers *et al.*, 2007; Frank, 2007; Boin, 2008). In addition to geospatial data quality, reusing data in semantic interoperability requires an evaluation of metadata quality. However, data producers typically provide no quality evaluation of geospatial metadata for the intended application, nor for potential applications. In order to facilitate data reuse, an evaluation of the quality of geospatial metadata with regards to the potential applications should be proposed. Some researchers in non-geospatial domain have studied the quality of metadata (Bruce and Hillmann, 2004; Ochoa and Duval, 2006). Although these approaches do not take into account the application in which data may be re-used, one may apply the same method to a different usage. In addition, they provide useful concepts for geospatial metadata. Nevertheless, there is still no specific work on the external quality of geospatial metadata.

In this section, we propose a restricted set of indicators and a quantitative approach for evaluating the external quality of geospatial metadata with regard to the context of re-users in a semantic interoperability process. These indicators are grouped into two categories: *global indicators* and *metadata-specific indicators.* In the first category there are three indicators – *convenience of language*, *completeness*, and *trust* – that influence the overall metadata. The second category includes one indicator that affects the quality of each metadata element, i.e. *freshness*. For each indicator, the quality value is within the interval [0, 1]. The value 1 indicates perfect quality while the value 0 indicates completely poor quality.

3.1 Quantitative evaluation of metadata

3.1.1 Global indicators category

Convenience of language. This indicates the usability of a given language by those who must express and use geospatial metadata. We consider that a convenient language allows users to focus on metadata expressions, not on learning a new language. This indicator is measured according to the expressivity and the ease of understanding the language. In order to evaluate the convenience of a language, we propose a matrix (see Table 1) that can be used by data producers taking jointly into account both the expressivity of the language and the level of a user's knowledge of a language. For expressivity, producers can use an evaluation like the one proposed by Salehi *et al.* (2007). Then, the level of a user's knowledge of a language can be evaluated by consulting potential re-users. Knowing the expressivity and levels of user knowledge, producers can evaluate the convenience of the language using, for example, the matrix presented in Table 1. The values of this matrix are calculated by considering 'high' as 1, 'low' as 0 and 'medium' as 0.5. The average of intersecting columns and rows are calculated for each cell (e.g., high expressivity and medium knowledge result in (1+0.5)/2 = 0.7).

Table 1. Convenience of languages evaluation with respect to a user's knowledge level and language's expressivity.

		Level of user's knowledge of the language		
		High	**Medium**	**Low**
Expressivity	**High**	1	0.7	0.5
	Medium	0.7	0.5	0.2
	Low	0.5	0.2	0

In the example in Section 2.2, metadata is represented by a free natural language that has a medium expressivity when avoiding technical jargon. We suppose that the level of receiver's knowledge is high with such language. According to Table 2, the convenience of this language is 0.7.

Completeness. This indicator shows the quantity of available metadata elements with regards to the required elements. We recognize thematic, spatial, and temporal completeness. The evaluation of the completeness *P* is calculated as:

$$P = \begin{cases} w \times \dfrac{N_{Elt}}{N_{ReqElm}} & ; if \quad N_{Elt} < N_{ReqElm} \\ 1 & ; Otherwise \end{cases}$$

Where N_{Elt} is the number of thematic, spatial, or temporal metadata elements available, N_{ReqElm} is the number of metadata elements required in a specific context, and *w* is a predefined weight for thematic, spatial, or temporal metadata elements. This weight indicates the importance of completeness of each type in the context of reuse. If the number of available metadata elements is equal to or greater than the required elements, metadata is complete and its value is set to 1. If not, the ratio of existing elements on required elements shows the degree of completeness.

The numbers of elements N_{ReqElm} and the weights can be predefined by data producers by asking potential re-users about their required metadata elements. In the example presented in Section 2.2 (the road-network data), the emergency application's re-users need, besides the context elements available, another element specifying the type of the obstacle. Thus, if we suppose that the weight of spatial completeness is 1, then $P = 3/4$.

Trust. This indicator describes the degree of faith that we have in the provided metadata. A decrease of trust lessens the external quality of metadata. Generally, in a chain of interveners, such as in semantic interoperability, if information is transmitted in a sequential manner, the trust decreases with the number of interveners (Bédard, 1986; Moe and Smite, 2007). We evaluate the trust using the following function:

$$T = \frac{\sum \alpha_i}{N}$$

where α_i is the confidence given to the ith intervener. The value of confidence is between 0 and 1. *N* is the number of interveners that transmitted the metadata element. We consider that each intervener transmits a metadata element just once. In the provided example in Section 2.2, we suppose that the metadata of roads network went through three interveners in a semantic interoperability chain. The first

has a confidence of 0.8, the second has a confidence of 0.5 and the third has a confidence of 0.8. Consequently, $T = (0.8+0.5+0.8)/3 = 0.7$.

At this point, one must remember that these indicators do not aim at being complete or precise, but solely at indicating if potential problems may arise without having to make a complete analysis containing all details. Such a method is frequently used in complex domains involving large numbers of hard-to-measure factors, such as the fields of epidemiology, ecology, economics, etc. It has already been used in the evaluation of geospatial data quality by Devillers *et al*. (2007) and the proposed selection of indicators can be modified according to the context at hand if needed.

3.1.2 Metadata-specific indicators category

Freshness. This quality indicator shows the degree of freshness related to the use of a metadata element at a given time with regard to its lifetime. Accordingly, the value of freshness is determined by the age and lifetime of the metadata. In semantic interoperability, the age of the metadata element *e* is the difference between the time of interpretation $T_{int}(e)$ and the time of the definition of that context element $T_{def}(e)$. The freshness of the context element $F(e)$ is evaluated as follows:

$$F(e) = \begin{cases} 1 - \dfrac{T_{\text{int}}(e) - T_{def}(e)}{lifetime(e)} \; ; \; if \quad T_{\text{int}}(e) - T_{def}(e) < lifetime(e) \\ 0 \quad ; \qquad Otherwise \end{cases}$$

where *lifetime* is the expected period of time after which a metadata element is no longer meaningful. The freshness value of context element $F(e)$ decreases as its age increases. A low value of freshness has a negative impact on the quality of metadata element. Lifetime and the time of definition T_{def} can be introduced by a data producer. In the road-network data example (Section 2.2), we suppose that the metadata element "bold lines indicate restrictions for vehicles" was defined in the year 2000 when the dataset was created, and will be obsolete after 30 years. Then $F(e) = 1- (2008-2000/30) = 0.73$.

3.2 A Framework for evaluating the external quality of metadata

In this section, we present a framework to evaluate the external quality of metadata. As shown in Figure 2, this framework consists of two layers: the detailed layer and the global layer. The detailed layer is devoted to evaluate the quality of metadata elements based on the freshness indicator. The global layer consists of two parts. The first part is derived from the detailed layer by calculating the weighted average of the qualities of metadata elements. The second part presents the quality evaluation of overall metadata based on the convenience, completeness and trust indicators. Such global and detailed representation allows re-users to get a global picture (quality of overall geospatial metadata), and then dig into the quality evaluation of each metadata element when needed to have a detailed perspective. The approach of considering different layers for quality analysis bears similarities with the method proposed by Devillers *et al*. (2007).

In each layer, while good metadata quality indicates it is less likely to have a risk of data misinterpretation, a poor quality indicates a higher risk. In order to help re-users to intuitively understand the degree of such a risk, we provide a set of

warnings which are based on standard danger symbols proposed by ANSI Z535.4 (1991) and previously used in the geospatial domain by Levesque *et al*. (2007). These symbols show the degree of the risk of data misinterpretation related to metadata quality at each layer, and thereby stimulate appropriate responses to such a risk. For example, if the warning is 'Danger', it would be better not to reuse geospatial data by means of semantic interoperability. Table 2 shows an example of how warnings can be predefined with regards to quantitative values of geospatial metadata quality.

Table 2. An example of indicators based on the geospatial metadata quality.

Quality metadata	Indicator of risk of data misinterpretation
Q <0.2	DANGER
0.2 < = Q <0.5	WARNING
0.5 < = Q <0.75	CAUTION
Q >= 0.75	NOTICE

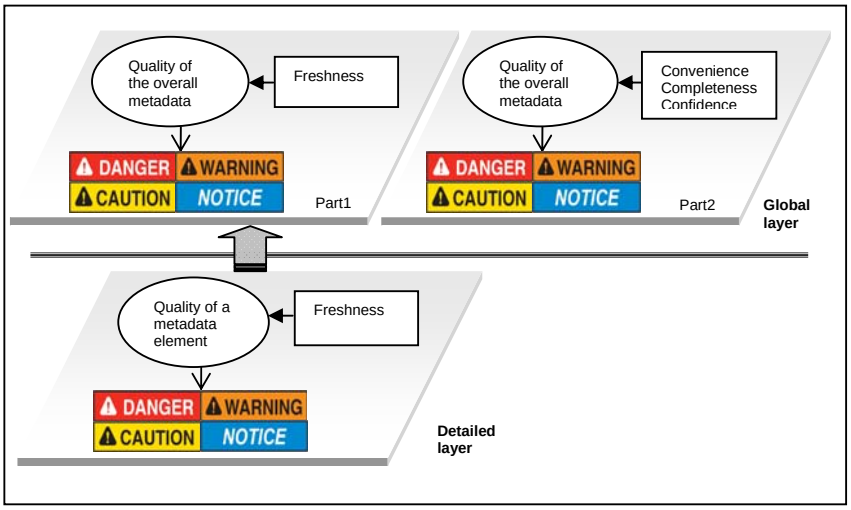

Figure 2. A framework for evaluating the external quality of metadata.

Referring to our example, in the detailed layer, the freshness of metadata element "bold lines indicate restrictions for vehicles" is 0.73; thus, for this element the 'caution' indicator has to be shown to the re-user. In order to calculate the freshness in the part one of the global level, we suppose that for two other elements of the metadata the freshness is 0.5 and 0.1. Thus, if the weight of these three elements is 1, the freshness of the overall metadata is the average of three values 0.73, 0.5, and 0.1, i.e., 0.44. The 'warning' indicator has to be shown to the re-user to indicate the risk associated to this part of global level. In the second part of the global level, the completeness is 0.75; thus, the 'notice' indicator has to be displayed.

4 Conclusion

Reusing geospatial data by means of semantic interoperability faces a risk of misinterpretation. This risk may be due to the fact that geospatial metadata, which is appropriate for the intended use, may be of poor external quality for data reuse. Evaluating this quality can help protect re-users from the risk of misinterpreting geospatial data. In this paper, we proposed a method to evaluate the external quality of geospatial metadata with respect to data re-users. For that, we introduced a set of indicators and quantitatively evaluated the quality of geospatial metadata. These indicators are organised into a framework consisting of two layers: the detailed layer and global layer. This framework allows re-users to get a global picture, and then drill down for the evaluation of the quality of each geospatial metadata element. In addition, we proposed a set of warnings that inform re-users about the risk of data misinterpretation related to metadata quality.

We should note that these indicators do not aim at completely eliminating the risk of data misinterpretation and reuse in the context of semantic interoperability, but rather represent a step forward in making re-users aware of such a risk. Further research is underway to define additional indicators such as relevancy and granularity of metadata elements. Then, the proposed approach will be implemented to show how the proposed indicators enhance semantic interoperability of geospatial data.

Acknowledgments

The authors wish to thank for its support the NSERC Industrial Research Chair of Geospatial Database for Decision Support financed by the Natural Sciences and Engineering Research Council of Canada, Laval University, Hydro-Québec, Research and Development Defence Canada, Natural Resources Canada, Ministère des Transports du Québec, KHEOPS Technologies, Intélec Géomatique, Syntell, Holonics, and DVP-GS. Also, we would like to thank the reviewers for their valuable comments.

References

Agumya, A. and. G.J. Hunter. (2002), "Responding to the consequences of uncertainty in geographical data". *International Journal of Geographical Information Science*, Vol 16(5):405-417.

ANSI (1991), *American National Standard for Product safety signs and labels*, ANSI Z535.4.

Bédard, Y. (1986), *A Study of the Nature of Data Using a Communication-based Conceptual Framework of Land Information*. PhD thesis, University of Maine, USA, 260p.

Bishr, Y. (1998), "Overcoming the semantic and other barriers to GIS interoperability", *International Journal of Geographical Information Science*, Vol. 12(4):299-314.

Boin, A. (2008), *Exposing Uncertainty: Communicating spatial data quality via the Internet*. PhD thesis, Department of Geomatics, The University of Melbourne, Australia, 183p.

Brodeur, J. (2004), *Interopérabilité des données géospatiales: élaboration du concept de proximité géosémantique*. PhD thesis, University of Laval, Canada, 267p.

Bruce, T.R. and D.I. Hillmann. (2004), "The continuum of metadata quality: defining, expressing, exploiting". *In:* D.I. Hillman and E.L. Westbrooks (eds.) *Metadata in Practice*, American Library Association, Chicago, US, pp. 238-256.

Devillers, R., Y. Bédard, R. Jeansoulin. and B. Moulin. (2007), "Towards spatial data quality information analysis tools for experts assessing the fitness for use of spatial data". *International Journal of Geographical Information Science*, Vol. 21(3):261-282.

Fonseca, F., M. Egenhofer, C. Davis, and K. Borges. (2000), "Ontologies and Knowledge Sharing in Urban GIS". *Computers, Environment, and Urban Systems*, Vol 24(3):232-251.

Frank, A.U. (2007), "Assessing the quality of data with a decision model". *In:* M. Molenaar, W. Kainz and A. Stein (eds.) *Modelling Qualities in Space and Time. Proceedings of the 5th International Symposium on Spatial Data Quality '07, ITC, Enschede, The Netherlands*. http://www.itc.nl/ISSDQ2007/proceedings/index.html.

Frank, A.U., E. Grum and B. Vasseur. (2004), "Procedure to select the best dataset for a task". *In:* M.J. Egenhofer, C. Freksa and H.J. Miller (eds.) GIScience 2004. LNCS, Vol. 3234: 81-93.

Harvey, F., W. Kuhn, H. Pundt, Y. Bishr and C. Riedemann. (1999), "Semantic Interoperability: A Central Issue for Sharing Geographic Information". *The Annals of Regional Science (Special Issue on Geo-geospatial Data Sharing and Standardization)*, Vol. 33(2):213-232.

Kuhn, W. (2005), "Geospatial Semantics: Why, of What, and How". *Journal on Data Semantics (Special Issue on Semantic-based GIS)*, LNCS, Vol. 3534:1-24.

Levesque, M.A., Y. Bédard, M. Gervais and R. Devillers. (2007), "Towards managing the risks of data misuse for spatial datacubes". *In:* M. Molenaar, W. Kainz and A. Stein (eds.) *Modelling Qualities in Space and Time. Proceedings of the 5th International Symposium on Spatial Data Quality '07, ITC, Enschede, The Netherlands.* http://www.itc.nl/ISSDQ2007/proceedings/index.html.

Moe, N.B. and D. Smite. (2007), "Understanding Lacking Trust in Global Software Teams: A Multi-Case Study". *In:* J. Münch and P. Abrahamsson (eds.) *Product-Focused Software Process Improvement*, LNCS, Vol. 4589:20-32.

Ochoa, X. and E. Duval. (2006), "Towards automatic evaluation of learning object metadata quality". *In:* J.F. Roddick, R. Benjamins, S. Si-Said Cherfi, R. Chiang, R. Elmasri, H. Han, M. Hepp, M. Lystras, V. Misic, G. Poels, I.-Y. Song, J. Trujillo and C. Vangenot (eds.) *Advances in Conceptual Modeling Theory and Practice*, LNCS, Vol. 4231:372-381.

Salehi M., Y. Bédard, M.A. Mostafavi and J. Brodeur. (2007), "On Languages for the Specification of Integrity Constraints in Spatial Conceptual Models". *In:* J.-L. Hainaut, E.A. Rundensteiner, M. Kirchberg, M. Bertolotto, M. Brochhausen, P. Chen, S. Sisaid Cherfi, M. Doerr, H. Han, S. Hartmann, J. Parsons, G. Poels, C. Rolland, J. Trujillo, E. Yu and E. Zimlanyi (eds.) *Advances in Conceptual Modeling – Foundations and Applications*, LNCS, Vol. 4802:388-397.

Schramm, W. (1971), "How Communication Works". *In*: J.A DeVito (ed.) *Communication: Concepts and Processes*. Prentice-Hall Inc., Englewood Cliffs, New Jersey, pp. 12-21.

Shannon, C.E. (1948), "A Mathematical Theory of Communication". *The Bell System Technical Journal*, Vol. 27:379-423.

Staub, P., H.R. Gnagi and A. Morf. (2008), "Semantic Interoperability through the Definition of Conceptual Model Transformations". *Transactions in GIS*, Vol. 12(2):193-207.