

# On The Problem Of Identifying The Quality Of Geographic Metadata <sup>?</sup>

Rafael Tolosana-Calasanz, José A. Álvarez-Robles, Javier Lacasta, Javier Nogueras-Iso, Pedro R. Muro-Medrano, and F. Javier Zarazaga-Soria

Computer Science and Systems Engineering Department,  
University of Zaragoza  
María de Luna, 1 50018 Zaragoza Spain  
rfael t, jantonio, jlacasta, jnog, prmuro, javyg@uni zar. es

**Abstract.** Geographic metadata quality is one of the most important aspects on the performance of Geographic Digital Libraries. After reviewing previous attempts outside the geographic domain, this paper presents early results from a series of experiments for the development of a quantitative method for quality assessment. The methodology is developed through two phases. Firstly, a list of geographic quality criteria is compiled from several experts of the area. Secondly, a statistical analysis (by developing a Principal Component Analysis) of a selection of geographic metadata record sets is performed in order to discover the features which correlate with good geographic metadata.

## 1 Introduction

Geographic Digital Libraries typically use geospatial metadata in order to provide surrogate representations of geographic resources and they represent the most powerful technique currently available for describing and locating geographic objects. As research and development make progress in the geographic area and metadata repositories grow in size (there are currently geospatial repository projects operating, whilst others are either to receive geographic metadata or plan to receive them in the near future), new requirements arise and system performance must improve necessarily. In this sense, the issues surrounding the creation of good quality metadata for Geographic Digital Libraries have surprisingly received little attention. Besides, regarding computer systems, there is a popular acronym, GIGO (Garbage In, Garbage Out), which means that if the input data is wrong, the output data will be unavoidably inaccurate or wrong.

---

<sup>?</sup> This work has been partially supported by the Spanish Ministry of Education and Science through the project TIC2003-09365-C02-01 from the National Plan for Scientific Research, Development and Technology Innovation. J. Lacasta's work has been partially supported by a grant (ref. B139/2003) from the Aragón's Government. Special thanks should be given to A. Sánchez, M. A. Manso, C. Fernández, P. Pachón, S. Fontano, A. Amaro and S. Muñoz

In other words, low quality information leads to bad system performance. Consequently, Geographic Digital Libraries need good quality metadata records in order to produce good results. The influence of poor quality metadata on the performance of Digital Libraries has been already studied from the perspective of other domains of knowledge: Barton [1] warned that "... these problems manifest themselves in various ways, including poor recall, poor precision, inconsistency of search results, ambiguities and so on. . .". Regarding the geographic domain, not only must the attention be focused on those problems, but also on the new ones that may appear with the geospatial information specific aspects: geographic coordinates, place names and so on.

Nevertheless, in order to tackle the problem, the requirements surrounding good quality metadata and, speaking more generally, the idea of quality have to be analysed previously. Quality is a matter of human judgement, thus, many complex human factors have a great influence on it. Additionally, it should be taken into consideration that these factors might vary widely among individuals or, what complicates things more, some individuals may modify their judgements throughout the time. However, the "notion" of quality is so simple, immediate and direct that it might be recognised less often by logical argument than by direct perception and observation. Mainly because of these reasons, much of the scientific research agrees that the definition of metadata quality is not out of difficulties. Nonetheless, according to [2] a metadata record of good quality is defined as "a record that is useful in a number of different contexts, both with respect to the search strategies and terms that can be used to locate it". Another definition [3], even more simple, might be "fitness for purpose". Following with this rationale, it seems that geographic metadata may be fit to their purpose, if they describe geographic data well and those descriptions are useful for their users.

The objective of this paper is to propose a quantitative method for quality assessment of metadata in geographic digital libraries. The method is developed through two phases, involving human experts in geographic information systems. Firstly, a list of geographic quality criteria, structural and semantic, is compiled from the experts. Then, derived from this criteria list, a group of metrics is proposed. Secondly, a statistical analysis of a selection of geographic metadata record sets is performed in order to discover the features which correlate significantly with good geographic metadata.

The remainder of this paper is organised as follows. Next section discusses other work related to this paper. In section 3, some geographic quality criteria are obtained from an opinion poll conducted to some experts and some geographic metadata metrics are proposed. In section 4, the statistical analysis is described and tested. Finally, the conclusions are given.

## 2 Related work

Initial efforts in metadata development have been primarily invested in structure rather than in content, that is, in the design and in the implementation of

geographic standards. Consequently, appropriate standards such as CSDGM [4] and ISO19115 [5] were developed and currently represent an excellent base for metadata creation and system interoperability. However, not only does metadata quality depend on these standards, but also on the creation process. Thus, generally speaking, two main approaches can be found in the research of metadata quality.

On the one hand, some studies are more concerned with the content of the metadata fields and the process involved in the creation of the metadata. In [1] it is stated that once a metadata standard has been implemented within a system, the specified fields must be filled out with real data about real resources and this process brings its own problems. The following assumptions underlying the metadata creation process in the learning objects and the e-Prints communities are also challenged there:

- { in the context of the culture of the Internet, mediation by controlling authorities is detrimental and undesirable, that rigorous metadata creation is too time-consuming and costly, a barrier in an area where the supposed benefits include savings in time, effort and cost.
- { only authors and/or users of resources have the necessary knowledge or expertise to create metadata that will be meaningful to their colleges
- { given a standard metadata structure, metadata content can be generated or resolved by machine.

Guy [3] suggests a number of quality assurance procedures that people setting up an e-Print archive can use to improve the quality of their metadata. The process is developed in the conviction that the metadata creation process is crucial to the establishment of a successful archive. Another interesting document is the report elaborated by the Academic ADL Co-Lab [2], which sets up the first step towards community creation and building in the learning repositories community. The paper is a guide to the various issues challenging learning repository projects: issues of quality, both content and metadata (creating quality content and metadata, guidelines to ensure access to quality educational content, quality and consistency of metadata, tools and workflow).

On the other hand, there exists another block of strategies whose research is mainly concerned with identifying and computing metrics for quality indicators. Then, resources are classified into different quality bands in accordance with those indicators. The study carried out by Armento [6] predicts quality rated Web documents (around popular entertainment topics) by using some pre-existing relevance ranking algorithms. Armento states that the results, though promising, should be tested more extensively and with more quantity of data in other knowledge domains. Other experiments carried out by Custard and Summer [7] identify and compute metrics for sixteen quality indicators (indicators that were obtained from an extensive and previous literature review and meta-analysis) and employ machine-learning techniques in order to classify educational resources into different quality bands based on these indicators. Additionally, previous experiments were developed to determine whether these indicators could be actually used for the classification. Hughes [8] describes the motivation,

design and implementation of an architecture to aid metadata quality assessment in the Open Archives Language (OAL) Community. It is worth highlighting that these quality indicators used in order to support quality judgements are based on the adherence to best practice guidelines for the use of the Dublin Core [9] elements and codes. Finally, another interesting work [10] computes some metrics for quality indicators and studies the relation between metadata quality and the quality of services.

### 3 Identifying Geographic Metadata Quality Criteria

At an early stage of our work, we considered studying the criteria by which the quality of geographic metadata records can be analysed. We carried out an initial experiment which consisted in asking several experts about the features, the elements or even the requirements for geographic metadata records that can determine their quality. As an outcome of this study, a compilation of geographic metadata quality criteria was obtained (see Fig. 1).

Two main tendencies can be observed in the compiled list. One tendency is more concerned with the structure of the metadata records and tries to determine to what extent the metadata records accomplish the standard. For instance, in the ISO 19115 standard, there exist certain recommendations regarding the format of certain data types such as dates, integers and so on. Additionally, in the same standard, there is a subset of elements known as the "ISO19115 Core metadata for geographic datasets" (called ISO Core onwards) that suggests to have each of them filled in. The most important elements such as the title, the abstract and the spatial reference system, among others, are included there. In the same sense, several experts were expecting to find specific information elements which were useful for their daily work and which were outside that core. Other considerations pointed out that the greater the number of filled elements, the higher the quality for the metadata record.

The other kind of tendency is related to semantic issues on the metadata elements. It is worth mentioning the considerations that the experts made on important free text elements such as the title and the abstract. Some experts stated that every title should answer, at least, the questions where, when, what and whom about the data; and that the abstract should describe, in a slightly broader way, the information which appears on the title, though they also thought that other issues can also be summarised there. Controlled elements such as the subject were found important as well, since they contribute to sort out subsets of topic-related records. The use of standardised thesauri, as the tool for filling in the subject element, was suggested as better than controlled lists. The rest of the semantic criteria focus their attention on more general aspects such as the coherence between the element and the information which it contains, the avoidance of duplicated information, the avoidance of contradictory information, the importance of precise information, the importance of homogeneity in the information among the metadata record set, and in a similar sense, the need for entity naming uniformity throughout the metadata record set and, finally, nat-

ural language semantic issues such as ambiguity which the experts recommend to minimise.

**Fig. 1.** Compilation of criteria for the assessment of geographic metadata quality

Nonetheless, some other interesting criteria taxonomies can be proposed:

- { according to the information type contained in the elements, the criteria may be sorted out into spatial (if they are related to spatial element types), textual (if related to textual element types) or temporal (if they deal with temporal element types).
- { assuming that geographic metadata records do not usually appear in an isolated way, but form geographic thematic catalogues whose topics are diverse, from environmental aspects, to geographic images and cartography maps, there may exist quality criteria related to individual quality aspects, global quality aspects and both of them. In fact, it seems obvious that the quality of the individual records affects the perception on the repository. For instance, let us consider a metadata record set in which a high percentage of the records does not present an important, desired characteristic (i.e. presenting an accurate title field, presenting a correct topic-keyword classification and so on). Although some records fulfil the requirements, the overall impression on the set is likely to be of bad quality, circumstance that is confirmed because wrong records appear more frequently. Consequently, quality criteria which measure individual quality aspects, global quality aspects and both of them have to be taken into account.

Additionally, when studying the initial classification of the criteria (structural criteria and semantic criteria) more carefully, it can be stated that the semantic criteria merely determine the constituents of metadata without any regard to the

quantity of each ingredient: they consider qualitative aspects of the metadata. On the contrary, the structural criteria give evidence of aspects which involve the measurement of quantity or amount which can be computed automatically.

**Table 1.** Proposal of geographic metadata metrics

<i>Metric ID</i>	<i>Metric name</i>	<i>Metric description</i>
Met1	purpose	Data purpose filled in
Met2	coreFilledPercentage	Percentage of the ISO Core filled in
Met3	alternateTitle	Number of words in the alternate title
Met4	numberOfFilledElements	Number of filled in elements
Met5	dataAccessConstraints	Data access constraints filled in
Met6	distributionFormat	Distribution format filled in
Met7	referenceSystem	Spatial reference system filled in
Met8	abstract	Number of words in the abstract
Met9	dataUpdateFrequency	Data and update frequency of the data filled in
Met10	title	Number of words in the title
Met11	responsiblesData	Information about the data responsible filled in
Met12	quality	Information about the data quality report filled in
Met13	lineage	Information about the lineage of the data filled in
Met14	metadataCreator	Information about the metadata creator filled in

In each engineering discipline, counting and measuring play an important role, because when it is feasible to measure the things that are being studied and to express them in numbers, something is known about them. In addition, an important element in proving theories is provided by experiments, without measuring, experiments would be useless as an aid to natural scientists and engineers. After these considerations on the significance of measurement, it should be noted that there are important difficulties when measuring geographic metadata quality and, what is more, the engineering good practice of observing, counting and measuring regarding geographic metadata quality has so far been neglected. Undoubtedly, those quantitative criteria compiled (the structural criteria from Fig. 1) represent a good starting point in order to obtain metrics for geographic metadata quality. In Table 1, a list of 14 metrics for assessing quantitative aspects of geographic metadata quality is proposed. Some of the proposed metrics merely determine whether certain elements appear on the records (i.e. purpose, dataAccessConstraints or quality), others count the number of words per element (i.e. title, alternateTitle or abstract) and others try to determine the percentage of elements in the ISO Core that are filled in.

## 4 Analysing Geographic Metadata Quality Criteria

### 4.1 Methodology

With the aim of understanding the notion of geographic metadata quality, we decided to carry out another experiment which intended to discover the quan-

titative features which correlate significantly with good geographic metadata. Basically, the experiment consists of the following steps:

- { select a sample of geographic metadata record sets
- { ask the experts to assess the quality of the record sets with a numerical assessment
- { compute the proposed metrics for the selected record sets
- { analyse the correlation between the metrics and the assessments coming from the experts.

**Table 2.** The average value of the assessment per metadata record set.

Because of the aforementioned reasons, the experiment was focused on the quality of the set rather than on individuals. Thus, 30 geographic metadata record sets of diverse cardinality were selected in order to carry out this experiment. They were compiled from different institutions: the Spanish National Geographical Institute, the French National Geographical Institute, several Spanish regional governments, some European institutions (such as the Joint Research Center) and the US Geological Survey. Their topics were Spanish, French and European cartography, Spanish and French hydrology, European LANDSAT images and orthoimages and geologic maps from the USA. The metadata record sets were all conforming to ISO 19115 with the exception of those from the US which were in CSDGM and were translated into the ISO 19115 standard by using the crosswalk described in [11]. Several experts from relevant public European organisations were asked to collaborate. Besides, the career backgrounds of the experts were rather heterogeneous: geographic, librarian and technologic.

The precise instructions given for the assessment were to assign a number from 1 (the lowest quality) to 10 (the highest quality) for each of the thirty metadata record sets and to write down an optional description for each of the assessments and a mandatory overall list of the assessment criteria. A form was given away in order to facilitate the noting down of those three elements. Two human-readable formats for the records were provided, one in HTML and another one in XML. A browser was recommended to visualise the records in the first case and the metadata edition tool CatMDEdit [12] in the second one. It is important to note, however, that neither evaluation criteria nor assessment

recommendations were indicated to them. However, as geographic metadata represent the description of a particular geographic dataset and the dataset was not provided, the assessment was somehow constrained.

**Table 3.** The numeric values of the metrics computed

Once the results were compiled, the first necessary step for this statistical analysis was to obtain a unique assessment value per metadata record set. The assessments of the experts, however, differed slightly. The variation depended on the nature of the criteria chosen, since some of the experts were more concerned with structural aspects and others with semantic ones. An arithmetic average on the assessments was calculated in order to have a unique number per record set (see Table 2, note that again the values range from 1, the lowest quality, to 10, the highest quality).

The 14 metrics were computed for each of the 30 sets. The process consisted in computing the metrics for each of the records and then computing the average of those values to obtain the metric for the record set (see Table 3).

One way of studying the correlation of the metrics and the metadata record sets quality might be by determining the main source of variation in the metrics. This study was carried out by developing a Principal Component Analysis (PCA) [13]. The PCA is a mathematical procedure that transforms a number of variables into a smaller number of uncorrelated variables known as principal components (PCs). The first principal component (PC1) accounts for as much of the variability in the information as possible, and each succeeding component accounts for as much of the remaining variability as possible. The aim of this



procedure is to reduce the dimensionality of data and to identify new meaningful variables. The relationship between the metadata quality values, coming from the assessments of the experts, and the principal component scores, obtained from the metrics, were studied through correlation analysis.

Fig. 2. The relationship between the quality values and the PC1

## 4.2 Results

Only the first component extracted from the PCA, which explained 32.2% of the observed variance (eigenvalue = 4.5), was significantly correlated with the metadata quality values (assessments). This correlation was strong and negative ( $R = -0.85$ ) as Fig. 2 shows. The *factor loading* of the PCA reflects (Table 4) that this component (PC1) was significantly correlated with the metadata metrics: *coreFilledPercentage*, *numberOfFilledElements*, *distributionFormat*, *referenceSystem*, *responsablesData*, *lineage* and *cataloguersData*. The numerical values represent the correlation degree between the metrics and the PCs and the symbol \* represents that there exists significant correlation ( $p < 0.001$ ). Thus, it can be concluded that these metadata metrics could be used as indicators of geographic metadata quality. If the value of the metrics increases, the quality of the record set increases as well. Nevertheless, the rest of the metrics were not significantly correlated and, consequently it cannot be statistically determined whether they have influence on the quality.

The first two components obtained through the PCA (PC1 and PC2) were used to represent the record sets in two dimensions (see Fig. 3). Metadata record

sets were sorted into three groups according to the degree of their quality value degree (high quality, >7; medium quality, 5-7 and low quality, <5 ). The highest quality group appears associated to low values of PC1 and the lowest quality group with high values of this component.

**Table 4.** The PCA *Factor loading*

Metric	PC1	PC2
purpose	-0.158	-0.603*
coreFilledPercentage	-0.486*	-0.595*
alternateTitle	-0.409	0.368
numberOfFilledElements	-0.794*	-0.221
dataAccessConstraints	-0.395	0.169
distributionFormat	-0.840*	0.015
referenceSystem	-0.710*	0.439*
abstract	-0.150	-0.781*
dataUpdateFrequency	0.441	-0.470*
title	0.424	-0.383
responsiblesData	-0.632*	0.269
quality	-0.402	-0.123
lineage	-0.616*	-0.305
metadataCreator	-0.853*	-0.193

According to Fig. 3, it is important to note that:

- { high quality metadata record sets appear quite near among them and far way from poor quality metadata record sets
- { high quality metadata record sets and some medium quality metadata record sets appear near what may suggest that the significantly correlated metrics do not determine quality completely and some other indicators such as those with semantic dimension take also an important role.

It can be stated that within this metadata set sample, the quality of the sets can be predicted by computing the correlated metrics. Thus, high values of the metrics involves medium-high quality and low values of them, low quality.

## 5 Conclusions

This work has presented early results from a series of experiments on identifying the quality of geographic metadata. The paper has proposed a quantitative method for quality assessment. The method is developed in two phases. Firstly, a list of geographic quality criteria was compiled from an opinion poll conducted to several experts of the area. The criteria were primarily classified into structural and semantic, though some other taxonomies were also described. The structural criteria give evidence of certain aspects which involve the measurement of quantity or amount which can be computed automatically. Derived from those

**Fig. 3.** Distribution of the different metadata record sets in relation to the PC1 and PC2

criteria, a list of 14 geographic metadata metrics was proposed. Secondly, a statistical analysis was carried out on a selection of 30 geographic metadata record sets. The experiment, by developing a Principal Component analysis, studied the relationship between the 14 metrics, which were computed for each record set, and the assessments made by some experts. As a result, it was observed that some metrics could be used as indicators of geographic metadata quality and, within the selected 30 record sets, the geographic metadata quality could be predicted by computing those metrics: high values of the metrics involve medium-high quality and low values of them, low quality.

As further work and in order to validate these results and to generalise them, the experiments should be carried out with an extended metadata corpus. Additionally, it would be interesting to investigate whether metadata quality metrics can be applied to the development of more efficient information retrieval ranking algorithms. It is expected that quality metrics can play an important role in computing the relevance of the resource described.

## References

1. Barton, J., Currier, S., Hey, J.: Building Quality Assurance into Metadata Creation: an Analysis based on the Learning Objects and e-Prints Communities of Practice. In: Proceedings of the 2003 Dublin Core Conference: Supporting Communities of Discourse and Practice-Metadata Research and Applications. (2003) ISBN 0-9745303-0-1.

2. Holden, C.: From Local Challenges to a Global Community: Learning Repositories and the Global Learning Repositories Summit. The Academic ADL Co-Lab (2003) Version 1.0.
3. Guy, M., Powell, A., Day, M.: Improving the Quality of Metadata in Eprint Archives. *Ariadne Magazine* (38) (2004) <http://www.ariadne.ac.uk/>.
4. Federal Geographic Data Committee (FGDC): Content Standard for Digital Geospatial Metadata, version 2.0. Document FGDC-STD-001-1998. Technical report (1998)
5. International Organization for Standardization (ISO): Geographic information - Metadata. ISO 19115:2003 (2003)
6. Armento, B., Terveen, L., Hill, W.: Predicting expert quality ratings of Web documents. In: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval. Does "authority" mean quality?, Athens, Greece (2000) 296 { 303 ISBN 1-58113-226-.
7. Custard, M., Summer, T.: Using Machine Learning to Support Quality Judgements. *D-Lib Magazine* **11**(10) (2005) ISSN 1082-9873.
8. Hughes, B.: Metadata Quality Evaluation: Experience from the Open Language Archives Community. In: Proceedings of the 7th International Conference on Asian Digital Libraries (ICADL 2004). Number 3334, Lecture Notes on Computer Science. Springer-Verlag (2004) 320{329 ISBN 3-540-24030-6.
9. International Organization for Standardization (ISO): Information and documentation - The Dublin Core metadata element set. ISO 15836:2003 (2003)
10. Zhang, B., Gonçalves, M., Fox, E.: An OAI-Based Filtering Service for CITIDEL from NDLTD. In: Proceedings of the 6th International Conference on Asian Digital Libraries (IACDL 2003). Number 2911, Lecture Notes on Computer Science. Springer Verlag (2003) ISBN 3-540-20608-6, pp 590-601.
11. Nogueras-Iso, J., Zarazaga-Soria, F.J., Lacasta, J., Bojar, R., Muro-Medrano, P.R.: Metadata Standard Interoperability: Application in the Geographic Information Domain. *Computers, Environment and Urban Systems* **28**(6) (2004) 611{634
12. Nogueras-Iso, J., Zarazaga-Soria, F.J., Muro-Medrano, P.R.: Geographic Information Metadata for Spatial Data Infrastructures - Resources, Interoperability and Information Retrieval. Springer Verlag (2005) ISBN 3-540-24464-6.
13. Jolliffe, I.T.: Principal Component Analysis. 2nd edn. Springer Series in Statistics. Springer Verlag (2002)