geofocus

Revista Internacional de Ciencia y Tecnología de la Información Geográfica
International Review of Geographical Information Science and Technology

Manso Callejo, M. A. y Bernabé Poveda, M. A. (2009): "Metadatos implícitos de la información geográfica: caracterización del coste temporal y de los tipos y tasas de errores en la compilación manual", GeoFocus (Artículos), nº 9, p. 317-336. ISSN: 1578-5157

GEOGRAPHIC INFORMATION IMPLICIT METADATA: CHARACTERIZATION OF TEMPORAL COST AND ERROR TYPES AND RATES IN MANUAL COMPILATION
METADATOS IMPLÍCITOS DE LA INFORMACIÓN GEOGRÁFICA: CARACTERIZACIÓN DEL COSTE TEMPORAL Y DE LOS TIPOS Y TASAS DE ERRORES EN LA COMPILACIÓN MANUAL

MIGUEL ÁNGEL MANSO CALLEJO, MIGUEL ÁNGEL BERNABÉ POVEDA
Dep. de Ingeniería Topográfica y Cartografía. ETSI Topografía. Geodesia y Cartografía
Camino de la Arboleda, s/n. Campus SUR-UPM. 20031. Madrid. España
m.manso@upm.es, ma.bernabe@upm.es

ABSTRACT

Traditionally and on the basis of the advice of some cataloguing experts, geographic information (GI) metadata are created manually. However, others think that the process is costly and error prone, hence the advisability of automatic methods should be looked into. In this paper, the time taken up and the errors made in the manual compilation of implicit metadata are examined. The study is based on a survey with 18 collaborators, 7 GI categories used, 40 datasets and 20 different storage formats; it models the time taken up, identifies, sorts and quantifies the most common errors. The results confirm that the time taken up is reduced with experience. New results are also provided: namely, that the cataloguing time is longer for raster data than for vector data and that the compilation of geographic extent and spatial reference systems is most likely to cause errors.

Keywords: compilation, manual, metadata, implicit, cost, time, errors

1. Introduction

One traditional definition of the term *metadata* is "*data about data*", with the first references to this term appearing in the context of geographic information, in ANZLIC (1996) and Kildow (1996). If we look for the origins of the term *metadata*, we will find its roots in the Greek word "μετα", «beyond» and the word "data", the plural of the Latin term *datum-i*, «piece of information» (RAE). Therefore, the meaning of the word may be explained as "beyond data". However, according to Howe (1993), the term *metadata* did not appear in print until 1973, despite having been coined by Lack Myers in the 1960s in order to describe sets of data and products. In the literature related to this subject we find a good number of authors who provide the interpretation and scope of the practical and theoretical meaning of the term. Among these,

Recibido: 2/9/2009
Aceptada versión definitiva: 15/10/2009

we find Caplan (1995), Milstead, Feldman, Ercegovac (1999), Sheldon and Steinacker *et al.* (2001), Swick and Duval *et al.* (2002) or Woodley *et al.* (2003). Summing up the contributions of all these authors, we may define the term eclectically as *the structured set of data that describe other data and whose purpose is to improve our knowledge of the described information and help us answer such questions as 'what', 'who', 'where', 'when', 'how much' and 'how'*. They may also be described as those autonomous products that, linked to the data, allow us to keep an inventory of these, enabling its publication and reference value through the catalogues kept in infrastructures of spatial data and, finally, allowing for the reutilization of data. The importance of metadata has been recognized by entities such as the EU's INSPIRE[1] Directive, and also by the endorsements of the GSDI[2] initiative.

1.1 Metadata classification.

In order to achieve those characterizations identified as the goal of this article, we will adopt Jokela's (22001) classification of metadata into *implicit* and *explicit*, based on the links or connections of the metadata with the data.

a. Implicit metadata are those that can be obtained from the data themselves (number of rows, columns or type of compression of the data).
b. Explicit metadata are those that cannot be obtained from the data themselves, being described instead in a separate file with a view to their cataloguing (for example, the name of the format employed for storing purposes) (Balfanz, 2002). For Beard (1996) and Díaz *et al.* (2008), metadata can also be *inferred* or calculated from other metadata or from the data themselves (for example, a toponym inferred from the coordinates) and for Goodchild (2007), metadata are those that can be obtained by data mining techniques.

These and other authors provide various other classifications of metadata, which we do not find as strongly connected to the purposes of our article: for example, *static, dynamic, temporary, descriptive, structural, control-related, contextual, semantic* or *administrative.*

1.2 Metadata creation.

Metadata can be created by using various methodologies: (a) manually, (b) automatically, (c) semiautomatically and (d) mixed versions of all the former. Several authors and projects have reviewed these creation techniques and suggested different solutions. For example, some authors state that the automatic procedures of metadata creation are unable to provide the information that can be provided by both data producers and metadata compilers (Campbell, 2008; Guy *et al.*, 2004; JORUM, 2004); others point out that manually created metadata, whether performed by the author of the data or automatically, can not provide the cataloguing experience and skills of information management experts (Currier *et al*., 2004; Guy *et al.*, 2004; JORUM, 2004); and finally, some authors state that manual metadata creation is a time-consuming, error-prone process (Batcheller, 2008; Najar, 2006; Wyoming; West and Hess, 2002; Leiden *et al.*, 2001; Guptill, 1999).

1.3 The goals of our article.

[1] Infrastructure for Spatial Information in Europe. Web: http://inspire.jrc.ec.europa.eu/

[2] Global Spatial Data Infrastructure. Web: http://www.gsdi.org

Recibido: 2/9/2009
Aceptada versión definitiva: 15/10/2009

As a first goal, this article sets out to measure the time required to manually obtain and transcribe a set of implicit metadata stored along with the data, and thus to be able to evaluate from objective data the first part of the aforementioned sentence: "*manual metadata creation is a time-consuming (...) process*".

The second goal of this article is to identify, classify and measure the error rates that take place in the manual creation of metadata. This way we will obtain objective data that may confirm or contradict the second part of the statement: "*manual metadata creation is a (...) error-prone process."*

Those errors derived from the manual creation of metadata can be divided into: errors or mistakes, and lapses or slips (Norman, 1980). Other authors such as Maurino *et al.* (1995) and early Reason (1990), when trying to study analytically error in systems, have classified them according to standards related to human behaviour. These standards are based on models for the performance of functions such as skill-based models, rule-based models and knowledge-based models (Rasmussen, 1983). Thanks to psychological contributions we know that these errors are affected by cognitive factors, memory capacity, attention capacity and schemas acquired through experience (Moment, 2008). Sometimes, what we know or think (cognitive factors) is in conflict with programmed actions and leads to the appearance of errors. Similarly, the cognitive principles that rule the way we perceive and reason things may affect our decisions, which in turn provokes errors. The complexity of the procedures involved in the cataloguing of GI may lead to an overflow of the memory required for a given task, eventually leading to overload and degradation of the performance. Therefore, the design of procedures must be carried out in such a way that the operator is able to keep his/her attention – otherwise, errors will occur. Wickens and McCarley (2008) devote a chapter of their "*Applied attention theory*" to attention control as a means of reducing errors, thus implying the importance of the subject.

Two factors make us aware of the fact that human errors may occur in the manual creation of metadata for GI: the cataloguer's cognitive aspects (the notions required for the interpretation of those aspects that demand a previous knowledge of, for example, spatial reference systems or coordinate systems), and the memory capacity required when there is a significant amount of operations to be performed. Another aspect that must be taken into consideration when studying human errors is the operator's attention in routine procedures such as the systematic manual creation of metadata about the elements of a cartographic series. All these factors seem to be important and may affect error rates, which may consequently affect productivity.

The rest of the article will be structured as follows: in chapter 2 we will describe how the study has been designed, how the datasets were prepared, how we designed the survey in order to collect data for the study, how the guide document for collaborators was prepared and, finally, the criteria for selecting the group of collaborators who took part in the study. In chapter 3 we will describe how we processed the obtained information after having integrated it into a database, how we chose to present the results of the study (both the results related to the time invested in obtaining and transcribing metadata classifying them by the nature of the data, and the results employed to characterize error rates as typos, coordinate interpretation errors or spatial reference systems errors). In chapter 4 we will discuss the obtained results and finally present our conclusions.

2. Study

We have identified two main goals in our study: the measure of the time required to locate and transcribe a set of implicit metadata about GI, and the cataloguing and quantifying of human error rates in the process.

2.1. Design of the study

The first goal of this study is to quantitatively measure the temporal cost/effort required by an operator to manually collect certain metadata elements. GI is characterized by the variety of data types (images, tables, and vector data) and by the diversity of existing storing formats. For this study we selected the following types of GI: a) old paper maps, b) rasterized and georeferenced cartography, c) digital ortophotography, d) digital elevation models, e) multispectral images, f) vector cartography and g) vector layers in formats used by geographic information systems (GIS). As for the diversity of storing formats, we selected the following: *PNG, JPEG, TIF, GeoTIFF, ECW, MrSID, JP2, ERS, IMG, BIL, PIX, XYZ, DEM, ASC, DGN, DAT, DXF, SHP, E00 and ADF*.

We left out of the study the GI stored in spatial databases (*Oracle, DB2, SQL Server, MySQL, PostGIS*, etc.) for two reasons: first, the need of specific, complex tools in order to access them, and second, the connectivity difficulties that may appear in the various communication networks when their administrators set limits to the connectivity to certain ports used by these networks (i.e., 1521: Oracle and 5432: PostgreSQL).

2.2. Dimension of the study

In order to define the dimension of the present study, we considered the representativeness of the file formats in each selected type of GI as well as the amount of data to be dealt with by each collaborator in the study. Bearing in mind the results of previously conducted pilot schemes in the field of training (Manso and Bernabé, 2006), in which the main conclusions were: (1) the creation of the first piece of metadata takes up time periods ranging from 45 minutes to 3 hours, and the creation of second and third metadata of the same subject require less time and present a narrower range (6-12 minutes or 4-9 minutes for the second and the third respectively), (2) a significant percentage of compiled data are incorrect (5-10% displace the decimal divider, 2-7% mistake latitude for longitude and 5-10% do not interpret or transform coordinates correctly). For the present study, the number of data sets to be described by each collaborator was set at 40. This number was defined with the aim of sparing the collaborators an overload of work that may eventually affect the study's results in a negative way. The initial estimation of the time each collaborator would devote daily was 5 hours. They were advised to report in case the devoted time went above this threshold, so that a decision could be agreed in this respect. In table 1 we show the various selected formats (20) for the seven GI categories.

2.3. Provided information and contents of the survey

The collaborators in this study received, along with the data and a form with data to be filled in, (a) a guide document describing the contents of the works, (b) the information categories and the amount of these, (c) the name of the initial data and (d) the fields to be filled in. The guide document included the instructions and guidelines to be followed in case of doubt. The provided form consisted of a spreadsheet with the following columns:

- GI category,
- Identification of the data set,
- Time (in minutes),
- Name of the employed IT tool,
- Spatial reference system (code EPSG),
- Coordinates West, East, North and South,
- In case of images:
  - Number of bands
  - Number of rows
  - Number of columns
  - Size of the pixel
- In case of vector files:
  - Type/s of geometries,
  - Number of points,
  - Number of lines
  - Number of polygons.

2.4. Profiles of the participants in the study

One of the goals of this study is to quantify the time invested by a data cataloguer in the recollection and transcription of the metadata elements in order to make it as representative as possible. Therefore, we consider an important aspect of this study the selection of collaborators' profiles, bearing in mind that their knowledge of geographic information must be reliably close to reality, and that their usual field of knowledge must be related to geodata. After contacting several candidates and asking them to commit themselves to the task, we selected 18 individuals whose academic backgrounds can be classified in 10 types, and with varying degrees of knowledge on subjects such as GI, GI cataloguing or metadata creation. In table 2 we describe both the technical profile and the experience with GI metadata of each collaborator.

2.5. Processing of the data collected in the study

The data received from the collaborators in spreadsheets have been transferred to the tables of a database, classifying the results in two big groups of GI: a matrix group and a vector group. The received information has been complemented by adding a group of additional columns in which author, type of GI and recording reference are codified and identified in each GI category (the first one is identified numerically and textually; the second and the third, only numerically). This way, the query building about the tables (with various levels of granulation and aggregation) is made easier. In chart 1 we describe as class diagrams the tables on which the analysis was conducted.

In order to fulfil the first goal we characterized the time required by a human operator to obtain, with the help of IT tools, those information elements selected for the metadata, and to manually copy those values in a spreadsheet. This is the same sequence of operations that would be followed by any cataloguer in a real case: to seek the information first, and then to write it on a metadata editor.

In order to satisfy this first goal we have designed a set of queries about the tables of the database; this way, the results have given us a better knowledge of the time taken up or the difficulties encountered when carrying out the assigned task in each GI category. Next we will present a few of the conducted queries:

```
SELECT AVG([Time]) FROM raster;
SELECT AVG([Time]) FROM vector;
SELECT AVG([Time]) FROM raster WHERE Category=1;
SELECT AVG([Time]) FROM raster WHERE [Sequence-category]=1;
SELECT AVG([Time]) FROM raster WHERE Category =1 AND [Sequence-category]=1;
SELECT AVG([Time]) FROM vector WHERE Category =1;
SELECT AVG([Time]) FROM vector WHERE [Sequence-category]=1;
SELECT AVG([Time]) FROM vector WHERE Category =1 AND [Sequence-category]=1;
```

The first approximation we tried in order to fulfil this goal was obtained through the first two aggregated queries for every element in the table: average time period of all raster and vector metadata. Although these values have proved useful when gauging efforts and assigning budgets to the task of metadata creation, other results of the analysis may also help us improve the qualifying of the involved cataloguers, as they allow us to identify sources of errors and other deficiencies. Thus, for example, we may obtain other results from the analysis of the collected information:

a. Values grouped by categories,
b. Values grouped by the sequence in the capturing process,
c. The completeness of delivered data,
d. The accuracy of the various attributes,
e. The correction of the values captured as metadata elements and their interpretation,
f. Accuracy in the interpretation of the type of coordinates of the data.

The first calculations have been automated as queries conducted to the database just as it has been presented. In other cases such as analysis of errors and accuracy of the obtained data, the procedure is more complex and has required a manual and systematic revision of the data stored in tables. Every error type and imprecision was recorded in order to eventually calculate the statistics, which define the aggregated correction or accuracy of the metadata. This last set of analysis sets out to fulfil the second goal of this study: to characterize error types and rates derived from manual metadata creation.

3. Description and analysis of the results

Next we will present the results obtained from the analysis of the time invested in the manual compiling of those implicit metadata defined in the survey, along with a series of reflections on these metadata. Next, we will briefly describe an obstacle that in some cases made it difficult to carry out the assigned tasks – an obstacle related to the employed IT tools' access to data. Finally we will present the results of the identification of detected error types, the corresponding categories in which the results have been classified and the rates for each error category.

3.1. Temporal cost of manual metadata creation

We have calculated both the average value and the standard deviation of the time invested in obtaining and copying metadata elements, classifying the results by category and creation order for any given category. The results of this analysis are shown in table 3, and next we will show our interpretation of the results.

The average time period required for raster data (5,82 minutes) is perceptibly superior to the time required for vector data (3,92 minutes). This fact must be taken into consideration if the number of files to be catalogued through metadata is rather large.

The average time and standard deviation obtained from cataloguing maps on paper is considerably superior to the rest of categories. This shows that the times taken up by each differ greatly. When analyzing this category individually, we observe that the third map is a geological map, in whose legend we do not find the information that identifies the spatial reference system, and that its coordinate grid is expressed in projected coordinates. In some cases the collaborators have not completed the requested data, while in the rest of the cases we perceive a great difference between the group of those who have performed the calculations for coordinate conversion and those who have not and therefore have invested less time. (By the way, this also proves how necessary it is that cataloguers have a wide knowledge of cartography).

In order to interpret the main trend of the time invested in the process, we visualized graphically the average time periods of the first, second, third and fourth metadata created for each category. In charts 2 and 3 are shown the trend estimations of the time periods invested in different categories.

The general trend estimation of the times invested for the first element of each category in comparison with the following ones is a decreasing one, except for one case (digital elevation models), as can be seen in chart 2. This decreasing trend is justified by the decreasing curve of learning processes (Wright-Patterson, 1936) during the operator's acquisition of skills.

The same decreasing trend can be observed in the average time periods invested in the first metadata of each category as a consequence of the learning process. In our study we propose a template for the recollection of results - a template in which we have sequenced categories, consequently leading the cataloguer to follow this sequence.

The main factors that may affect this trend (3rd source paper maps and ortophotography, 3rd and 4th source for MDE) are: a change in the type of information to be described (the switch from a topographic map to a geological map, which requires a coordinate conversion), and those changes derived from the particularities of file formats involved in one category or from the amount of stored information (size of the files).

Next we will proceed to interpret individually the results of the study, analyzing those problems detected by the collaborators, and studying each information element to be compiled from the geographic information.

3.2. Difficulties in the access to data

One main difficulty encountered by the collaborators was the access to implicit metadata elements stored in PIX format (PCI). The lack of knowledge (or inexistence) of a software application capable of recognizing and processing the above mentioned format is the reason why in the 14% of the cases the survey was not complete.

3.3. Errors

We have detected errors in the following requested metadata elements: the number of bands, the counts of rows and columns, the sizes of the pixels, the geographic extents and spatial reference systems of the data. Next we will observe the results in detail with the help of a set of tables and charts.

3.3.1. Identification of bands

One essential, implicit piece of metadata in raster data is the number of bands. In some cases, those values linked to the pixels were stored in a spaghetti-type format, by bands, while in other cases they were interlaced. The analysis of this element shows that many collaborators were unable to obtain this piece of data, or assigned incorrect values to it. In table 4 we show the aggregated results of this analysis. 51,2% percent of the records left this element blank, while 12,9% of the data were incorrect and only 35,8% were correct.

3.3.2. Identification of Rows and Columns

Other implicit metadata we consider essential in order to deal with raster data are the number of rows and columns employed to identify a pixel in the grid. The analysis of these elements shows that in most cases they were correctly filled out. In some cases they were left blank, while others were incorrect and, in the rest, values had been permutated. Table 5 shows the results of the analysis of these data: in 89,3% of the cases they were correct, in 2,67% of the cases they had been left blank, in 2,6% of the cases we found typographical errors and finally, in 6,5% of the cases the count of rows had been mistaken for the count of columns.

3.3.3. Capture of the pixels' spatial resolution

Some errors detected when analyzing the resolution of the pixel in matrix data show a lack of knowledge – for example, the inclusion of this piece of information in cases in which it should not be included (for example, paper maps), with an error rate of 57%. In other cases (25%), it should have been identified but it has not; in 4% of the cases, the units in which the

Recibido: 2/9/2009
Aceptada versión definitiva: 15/10/2009

resolution is expressed were not identified, and finally in 1% of the cases the measurement unit was shown in dots per inch. On the other hand, we would like to highlight the fact that the measurement units described by the collaborators did not appear in an standardized way, adopting instead various values such as M, m, m/px, *meters*, *metres*, metros, etc.

3.3.4. Identification of types and amount of geometries

Small percentages of errors were detected in the identification of types of geometries (points, lines, polygons, etc.) (5,17%), and typographic errors occurred when copying the amounts of geometries (2,16%).

3.3.5. Capture of the geographic extent

Those errors detected in the analysis of the geographic extent can be classified into: non-captured coordinates, non-geographic coordinates, geographic coordinates in a complex format (degrees, minutes, and seconds) and lapses committed by forgetting the negative sign for negative longitudes. We have detected high percentages (71% and 54%) in which the collaborators did not fill out correctly the coordinates when writing them in a complex format. We must also highlight the low percentage of cases in which the coordinates were correct (10,5% and 5,1%). Table 4 shows the rates of each category listed for the groups of raster and vector data.

Systematic errors of this kind occurred due to the lack of a specific training in cataloguing, or the lack of explicit procedures for the development of those activities required to obtain the desired result.

3.3.6. Capture of the Spatial Reference System (SRS)

Those errors detected when analyzing the spatial reference system can be classified into: incorrect, non-captured and correct-but-not-codified-with-EPSG-identifiers. Table 5 shows numerically the rates of correct answers in the identification of the SRS, both in raster and vector data. We must point out that only between 25% and 28% of the metadata that identify the SRS did it correctly by using standardized identifiers, while an important percentage of them did not provide the EPSG identifier even when the identification was correct (34% and 46%) and finally, between 20% and 36% of the cases presented errors or simply had not obtained results.

These errors occurred due to: the lack of training in geodesy and reference systems, and the lack of a specific training in cataloguing methods in which procedures are systematized in order to obtain those EPSG identifiers linked to the SRS starting from the data obtained from IT tools.

4. Discussion and evaluation of the results

Recibido: 2/9/2009
Aceptada versión definitiva: 15/10/2009

The lack of studies devoted to the quantifying of time required by a GI cataloguer in order to obtain implicit metadata has been one of the reasons that led us to carry out the present study. Therefore, the contributions derived from this work are the following ones:

1. The statistical characterization of the temporal cost and its behaviour with the cataloguer's experience. This contribution is of unquestionable importance for those companies offering cataloguing services for geographic information, and for those geo-institutions that must foresee the costs derived from metadata creation.
2. The identification of various sources of errors when manually locating, interpreting and transcribing implicit metadata. This work solves this existing lack.
3. The detection of systematic errors derived from the lack of training or the lack of established procedures for cataloguing geographic information. This reveals the need of training experts in GI cataloguing – experts capable of fulfilling institutional demands.
4. The study also reveals the existing difficulty in identifying those spatial reference systems (SRS) employed in the formats and in expressing them through EPSG codes.

Therefore, this study's main contribution lies in the detection and statistical characterization of the error types that occur when manually capturing implicit metadata.

5. Conclusions

The results of the study show empirically and quantitatively that manual metadata creation is a costly, error-prone process. The aspects of the costs have been characterized as the time required by a cataloguer to obtain and transcribe implicit metadata and the most frequent error types, along with their ratios.

We have ascertained the effect of the learning curve in cost-related aspects, showing that the employed time tends to stabilize after the third or fourth metadata record in any given category.

The study has allowed us to detect several error types that take place when manually capturing the implicit metadata:

- Mistaking the count of rows for the count of columns,
- Omitting the negative sign in West longitudes,
- Expressing the geographic extent through projected coordinates instead of geographic coordinates,
- Expressing the longitudes and latitudes in a complex way instead of a decimal way.

We have detected some difficulties in accessing those implicit metadata elements in PIX format. The wide existing range of storing formats makes it difficult for one IT tool to access all of them, which consequently affects manual metadata creation.

The obtained results confirm the initial hypothesis: "manual metadata creation is a time-consuming, error-prone process", at least in the case of implicit metadata. Given these premises, it is advisable to further investigate methodologies of manual metadata creation and develop new tools capable of accessing many types of formats in order to extract implicit metadata, reducing the temporal cost of the cataloguer and avoiding different sources of errors.

6. Acknowledgements

We would like to acknowledge the collaborators of the query without whom the study would not have been possible: Alejandra, Ana, Carlos, César, Daniel, Daniela, Emilio, Esther, Javier, María, María del Mar, María José, Miguel, Paloma, Rafael and Sara.

The Instituto Geográfico Nacional contributed to the results described in this study through the UPM-P076005609 and UPM-P086005572 collaboration agreements.

Bibliographic references

ANZLIC (1996): "ANZLIC Guidelines: Core Metadata Elements Version 1 Report". *ANZLIC Working Group on Metadata*.

Batcheller, J. (2008): "Automating geospatial metadata generation — An integrated data management and documentation approach". *Computers & Geosciences*, 34, pp. 387–398.

Balfanz, D. (2002): "Automated Geodata Analysis and Metadata Generation". *Society of Photo-Optical Instrumentation Engineers -SPIE-, Bellingham/Wash. Visualization and Data Analysis 2002*. San Jose, USA Bellingham/Wash.

Beard, K. (1996): "A Structure for Organizing Metadata Collection". *Third International Conference/Workshop on Integrating GIS and Environmental Modeling*, Santa Fe, New Mexico, USA.

Campbell, T. (2008): "Fostering a Culture of Metadata Production". *GSDI10: Tenth International Conference for Spatial Data Infrastructure*, St. Augustine, Trinidad. Available at http://www.gsdi.org/gsdi10/papers/TS8.2paper.pdf

Caplan, P. (1995): "You call it corn, we call it syntax-independent metadata for document-like objects". *The Public Access Computer Systems Review*, 4, 6.

Currier, S., Barton, J. and others (2004): "Quality assurance for digital learning objects repositories: issues for the metadata creation process". ALT-F, *Research in Learning Technology*, 12, 1.

Díaz, L. Granell, C. Beltrán, A. Llaves, A. and Gould, M. (2008): "Extracción Semiautomática de Metadatos: Hacia los metadatos implícitos". *II Jornada de SIG Libre*. Universidad de Girona.

Duval, E., Hodgins, W., Sutton, S., and Weibel, S. (2002): "Metadata Principles and Practicalities", *D-Lib Magazine*. [Consulted on 10-1-2008]. Available at http://www.dlib.org/dlib/april02/weibel/04weibel.html.

Ercegovac, Z. (1999): "Introduction". *Journal of the American Society for Information Science*, 50, 13, pp. 1165-1168.

Goodchild, M. (2007): "Citizens as Voluntary Sensors: Spatial Data Infrastructure in the World of Web 2.0", *International Journal of Spatial Data Infrastructures Research*, 2007, 2, pp. 24-32.

Guptill, S. G. (1999): "Metadata and data catalogues", in Longley, P., Goodchild, M. F., Maguire, D. J., Rhind, D. W. (Ed.), *Geographical Information Systems*. Wiley, Chichester, pp. 677–692.

Guy, M., Powell, A. and Day, M. (2004): *Improving the Quality of Metadata in Eprint Archives*. [Consulta: 1-10-2008]. Available at http://www.ariadne.ac.uk/issue38/guy/

Howe, D. (1993): *Free on-line dictionary of computing*. Available at: http://foldoc.org/index.cgi?Metadata.

Recibido: 2/9/2009
Aceptada versión definitiva: 15/10/2009

Jokela, S. (2001): "Metadata enhanced content management in media companies", *Acta Polytechnica Scandinavica*, Ma 114. Finnish Academies of Technology, Helsinki. [Consulted on 1-1-2009]. Available at: http://lib.tkk.fi/Diss/2001/isbn9512256932/isbn9512256932.pdf

JORUM (2004): "The JISC Online Repository for [learning and teaching] Materials", *JORUM Scoping and Technical Appraisal Study,* Volume V: Metadata.

Kildow, M. (1996): "The value of Metadata (A NSDI report)". *US Fisheries and Wildlife Services*. [Consulted on 10-1-2008]. Available at http://www.r1.fws.gov/metadata/meta.html.

Leiden, K., Laughery, K.R., Keller, J., French, J., Warwick, W. and Wood, S.D. (2001): "A Review of Human Performance Models for the Prediction of Human Error". *National Aeronautics and Space Administration*, Moffett Field, CA, USA, 125pp

Manso, M.A. and Bernabé, M.A. (2006): "Metadatos: Extracción y derivación automática de atributos". *JIDEE06, Jornadas Técnicas de la IDE de España 2006*, Castellón 18th-20th of October, 2006.

Maurino, D. E., Reason, J., Johnston, N. and Lee, R. B. (1995): *Beyond Aviation Human Factors*. Burlington, VT: Ashgate Publishing Company.

Milstead, J. and Feldman, S. (1999): *Metadata: Cataloguing by any other name*. Volume available at line 23, 1, pp. 25-31. [Consulted on 10-1-2008]. Available at http://www.onlineinc.com/onlinemag/OL1999/milstead1.html.

Moment, S. L. (2008): "A Compact Introduction to Human Error", *University of Illinois Human Factors Division Proceedings.* [Consulted on 10-1-2008]. Available at http://www.humanfactors.illinois.edu/research/HumanElementArticles/CompactIntroToHuman Error

Najar, C. (2006): *A model-driven approach to management of integrated metadata – spatial data in the context of spatial data infrastructure*. PhD thesis Nr. 16474, Institute of Geodesy and Photogrammetry, Eidgenössische Technische Hochschule Zürich. Available at http://e-collection.ethbib.ethz.ch/show?type=diss&nr=16474

Norman, D.A. (1980): "Errors in Human Performance", University of California, San Diego, (Ed.) *Center for Human Information Processing Report Nº. 8004*.

Rasmussen, J. (1983): "Skill, Rules, and Knowledge; Signals, Signs, and Symbols, and Other Distinctions in Human Performance Models", *IEEE Transactions on Systems Man and Cybernetics*, 13, 3, pp. 257-266.

Reason, J. (1990): *Human Error*. New York: Cambridge Press.

Sheldon, T. (2001): Linktionary. Entry «Metadata». [Consulted on 5-29-2006].

Steinacker, A., Ghavam, A., and Steinmetz, R. (2001): "Metadata Standards for Web-Based Resources". *IEEE Multimedia* 8,1, pp. 70-76.

Swick, R. (2002): *Metadata Activity Statement*. [Consulted on 5-29-2006]. Available at http://www.w3.org/Metadata/Activity.html

West Jr. and Hess, T. (2002): "Metadata as a knowledge management tool: supporting intelligent agent and end user access to spatial data". *Decision Support Systems* 32, pp. 247–264

Wickens, C. and McCarley, J. (2008): *Applied Attention Theory*, Taylor and Francis/CRC Press, Boca Raton.

Woodley, M. S., Clement, G. and Winn, P. (2003): *DCMI Glossary*. Dublin Core Metadata Initiative: Making it easier to find information. [Consulted on 10-1-2008]. Available at http://dublincore.org/documents/2003/08/26/usageguide/glossary.shtml

Wright-Patterson, T.P. (1936): "Factors Affecting the Cost of Airplanes", *Journal of Aeronautical Sciences*, 3(4) pp. 122-128.

Wyoming (2003): *Metadata Resources at the University of Wyoming*. Available at http://www.uwyo.edu/wygisc/metadata/why.html

TABLES

Table 1. Categories, formats and description of the scope of analysis

Category	Format	Contents
Paper maps	PNG	Topographic Map of the Ferrol area
	TIF	Nautical Chart Port of La Guardia, scale 1:10.000
	JPEG	Geologic Map of Segovia, scale 1:50.000
	ECW	Topographic Map of Galicia, scale 1:50.000
Rasterized and georeferenced	JP2	Nautical Chart anchorage of Médano de Santiago, scale 1:30.000
	TIF	Topographic Map of La Coruña, scale 1:200.000
	JPG	Topographic Map of La Coruña, scale 1:400.000
	ERS	Topographic Map of the Northwest third, scale 1:800.000
	ECW	Topographic Map of La Coruña, scale 1:400.000
Digital ortophotography	SID	Colour Ortophotography, scale 1:10.000
	TIF	Black and White Ortophotography, scale 1:10.000
	JP2	Colour Ortophotography, scale 1:10.000
	ECW	Colour Ortophotography, scale 1:5.000
Digital elevation models	IMG	Digital terrain model
	BIL	Digital terrain model
	PIX	Digital terrain model
	XYZ	Digital terrain model
	DEM	Digital terrain model
	ASC	Digital terrain model
Multispectral images	IMG	Hyperspectral image with 55 bands
	PIX	Hyperspectral image with 6 bands
	ERS	Hyperspectral image with 4 bands
	PIX	Classified hyperspectral image with 8 bands
	PIX	One-band classified image
	ERS	Hyperspectral image with 7 bands
	IMG	Hyperspectral image with 7 bands
	ERS	One-band image
Vector maps	DGN	Topographic cartography in CAD format
	DGN	Topographic cartography in CAD format
	DAT	Vector cartography – only dots.
	DAT	Vector cartography – only lines.
	DXF	Vector cartography – only lines.
	DGN	Topographic cartography in CAD format
	DGN	Topographic cartography in CAD format
GIS vector layers	SHP	Line-type layer related to fishing reserves
	SHP	Polygon layer with isopluvial lines
	SHP	Polygon layer with isothermal lines
	SHP	Polygon layer with administrative limits of populated areas
	E00	Layer with land use programs
	ADF	Layer with political limits

Table 2. Description of the collaborators on the study

Nº Col.	Qualifications	Degree of knowledge
3	Topographic Engineering Students	Wide experience in metadata creation
4	Topographers pursuing higher studies	Various degrees of experience and knowledge of metadata
2	Geodesy and Cartography Engineers	No experience with metadata.
2	Doctorate students, Geographic Eng.	Wide knowledge of metadata.
1	Degree in Biology and MBA on GIS	No experience with metadata.
1	Degree in Environmental Science	Lack of knowledge of geographic information or metadata.
1	IT Technical Engineer	Lack of knowledge of geographic information or metadata.
1	IT Engineer	Some knowledge of metadata.
2	Government-employed topographers	Wide knowledge of metadata.
1	Geographic Engineer	Wide knowledge of metadata.

Table 3. Average times and standard deviation by categories and sequence of the metadata

Category	Average time (minutes)	Stand. deviation	Meas. 1st metadata	Dev 1st	Meas. 2nd metadata	Dev 2nd	Meas. 3rd metadata	Dev 3rd	Meas. 4th metadata	Dev 4th
Paper maps	9,7	7,68	11,4	10,28	6,9	3,66	11,3	7,98	9,23	7
Rasterized and georeferenced cartography	6,2	3,84	7,2	4,99	7,2	3,45	5,6	3,58	6,22	4,12
Digital ortophotography	4	2,24	4,3	2,93	4,5	2,18	3,5	1,54	3,6	2,1
MDE/MDT	4,6	2,93	4,7	2,58	3,7	2,10	5,3	3,45	7,41	4,12
Multispectral images	4,7	3,09	6,2	4,08	4,5	3,5	5,2	2,67	3,66	3,44
Average times for Raster	**5,82**	**4,61**	**6,77**	**6,15**	**5,50**	**3,25**	**6,24**	**5,21**	**6,02**	**4,15**
Vector cartography	3,9	2,58	5,6	3,64	4,4	2,85	4,2	2,63	3,81	2,94
GIS vector layers	3,4	2,79	3,6	2,47	2,9	1,40	2,6	1,54	2,72	1,48

Average times for Vector	**3,92**	**2,68**	**4,58**	**3,22**	**3,66**	**2,35**	**3,32**	**2,25**	**3,26**	**2,21**

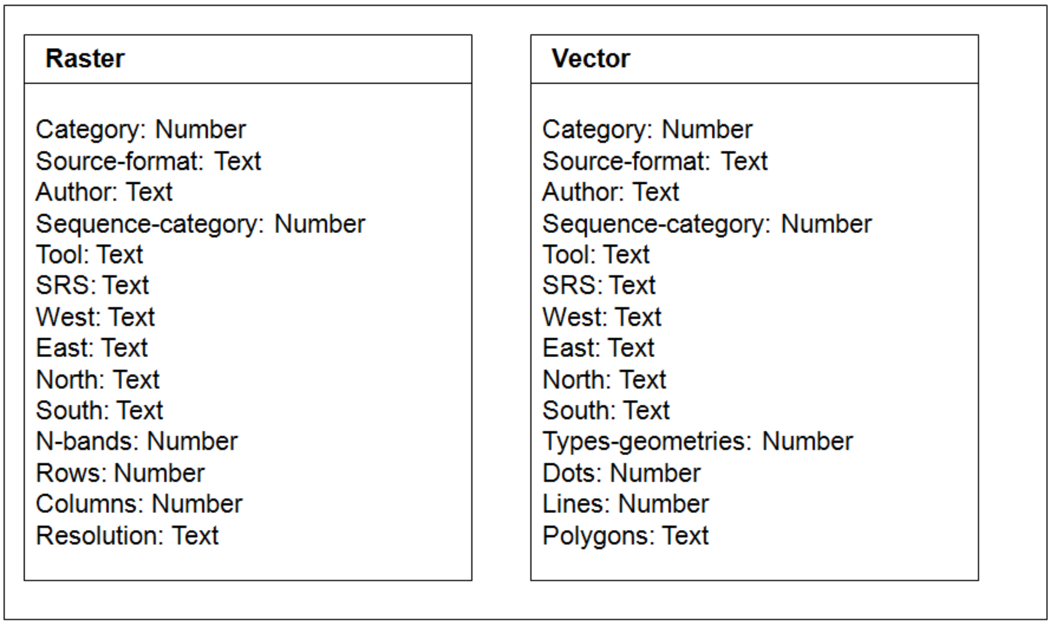
Table 4. Correct answer and error rates in geographic extent

Coordinates	**Raster**	**Vector**
Correct	10,5%	5,1%
Blank	18,72%	3%
Non-geographic (proj)	5,1%	8,2%
Complex format (º,',")	70,78%	54,3%
Error in West Longitude	29%	

Table 5. Correct answer and error rates in the identification of the SRS

Spatial Reference System	**Raster**	**Vector**
Correct (EPSG)	28,6%	24,5%
Correct, but not EPSG	33,9%	46,12%
Blank	8,8%	12%
Incorrect	28,6%	7,7%

CHARTS



Raster

Category: Number
Source-format: Text
Author: Text
Sequence-category: Number
Tool: Text
SRS: Text
West: Text
East: Text
North: Text
South: Text
N-bands: Number
Rows: Number
Columns: Number
Resolution: Text

Vector

Category: Number
Source-format: Text
Author: Text
Sequence-category: Number
Tool: Text
SRS: Text
West: Text
East: Text
North: Text
South: Text
Types-geometries: Number
Dots: Number
Lines: Number
Polygons: Text

Chart 1. Description of the tables employed to carry out the data analysis

Manso Callejo, M. A. y Bernabé Poveda, M. A. (2009): "Metadatos implícitos de la información geográfica: caracterización del coste temporal y de los tipos y tasas de errores en la compilación manual", GeoFocus (Artículos), nº 9, p. 317-336. ISSN: 1578-5157

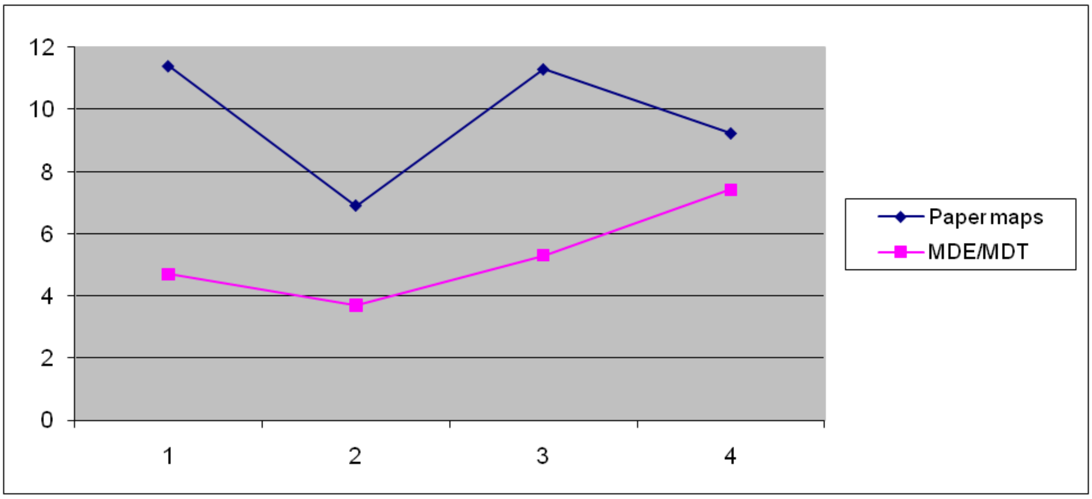


Chart 2. Trend estimation of the time invested in creating the first four metadata for the Paper maps and MDE categories

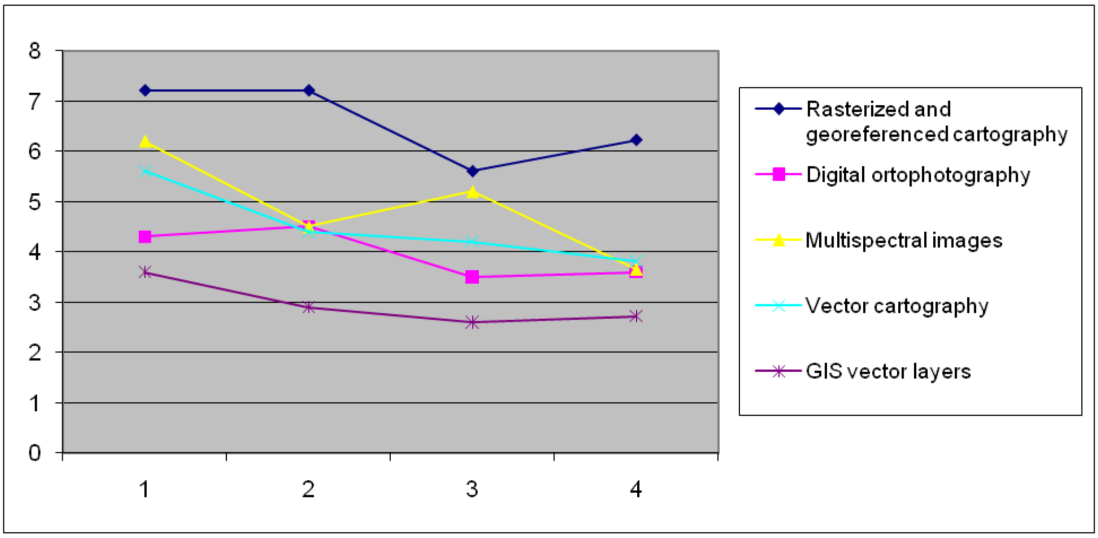


Chart 3. Trend estimation of the time invested in creating the first four metadata for the rest of the categories

Recibido: 2/9/2009
Aceptada versión definitiva: 15/10/2009

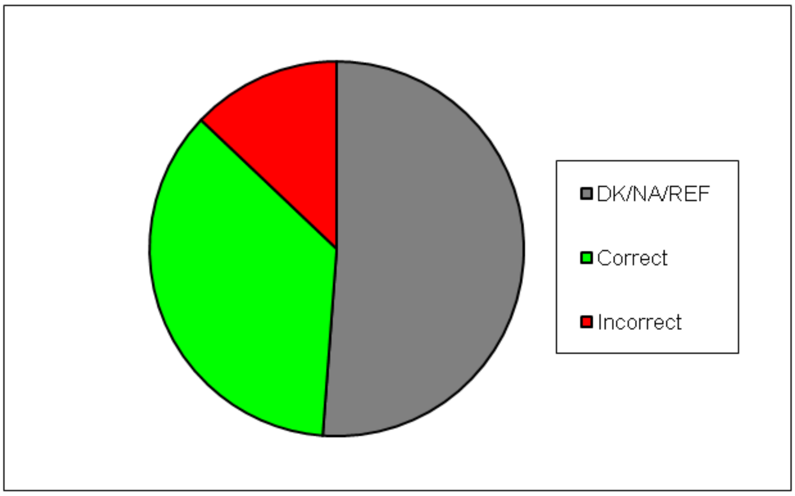


Chart 4. Correct answer and error rates in bands

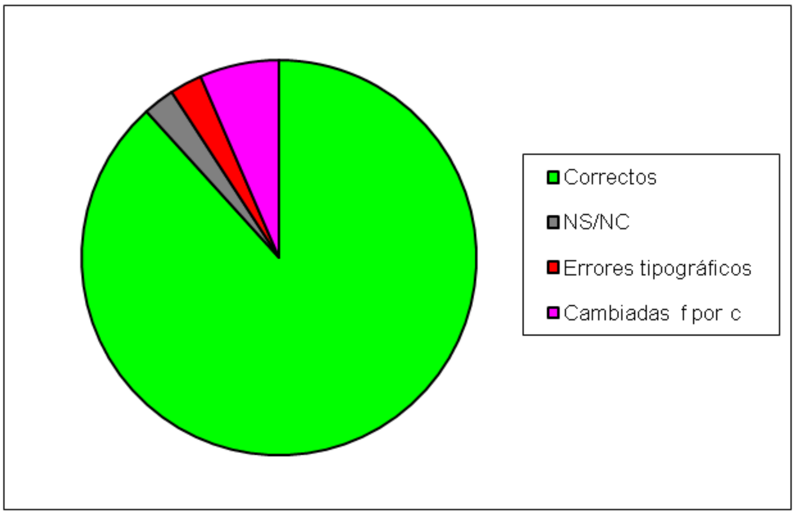


Chart 5. Correct answer and error rates in the count of rows and columns

Recibido: 2/9/2009
Aceptada versión definitiva: 15/10/2009