Sistemas y Tecnologías de Información

Actas de la 7ª Conferencia Ibérica
de Sistemas y Tecnologías de Información
Madrid, España
20 al 23 de Junio de 2012

Vol. II – Artículos

Editores
Álvaro Rocha
Jose A. Calvo-Manzano
Luís Paulo Reis
Manuel Pérez Cota

Artículos Cortos

Artículos Posters

Simposio Doctoral

aisti
Associação Ibérica de Sistemas e Tecnologias de Informação

POLITÉCNICA

Sistemas y Tecnologías de Información

Actas de la 7ª Conferencia Ibérica de Sistemas y Tecnologías de Información
Madrid, España
20 al 23 de Junio de 2012
AISTI | UPM

Vol. II – Artículos Cortos, Artículos Posters y Simposio Doctoral

Editores
Álvaro Rocha
Jose A. Calvo-Manzano
Luís Paulo Reis
Manuel Pérez Cota

ISBN:
978-989-96247-7-1

CRÉDITOS

TÍTULO
Sistemas y Tecnologías de Información

SUB-TÍTULO
Actas de la 7ª Conferencia Ibérica de Sistemas y Tecnologías de Información
Madrid, España
20 al 23 de Junio de 2012

Vol. II - Artículos Cortos, Artículos Posters y Simposio Doctoral

EDITORES
Álvaro Rocha, Universidade Fernando Pessoa
Jose A. Calvo-Manzano, Universidade Politécnica de Madrid
Luís Paulo Reis, Universidade do Minho
Manuel Pérez Cota, Universidad de Vigo

EDICIÓN, IMPRESIÓN Y ACABADO
APPACDM – Associação Portuguesa de Pais e Amigos do Cidadão Deficiente Mental, Braga, Portugal

DEPÓSITO LEGAL
?????

ISBN
978-989-96247-7-1

WEB
http://www.aisti.eu/cisti2012

Exploring alternatives for geodata preservation

Anita E. Locher, Miquel Termens
Dept. of Library and Information Science
University of Barcelona
Barcelona, Spain
alocher@ub.edu, termens@ub.edu

***Abstract*— We explore the activity of actors in geosciences and information science in relation to digital preservation of spatial data produced with public or private funds. The experience of four recent projects, two with the participation of libraries and two with archives, each in coalition with spatial data centres are compared. Their recommendation is applied to the context of the Institut Cartogràfic de Catalunya with the goal to develop a preservation strategy for spatial data produced and collected in Catalonia (Spain).**

Keywords- spatial data; geodata; digital preservation; GIS

I. INTRODUCTION

The priority for spatial data production centres is access to data on a short term, but they start to be interested in preservation because of user demand for time-based research. Easily accessible geospatial data are capitalized by an ever growing amount of new applications but the main argument for long term preservation in geosciences is longitudinal research. Climate and other kinds of environmental change analysis benefit from long term data preservation [1, 6, 7, 12]. Understanding change on our planet can help predict natural disasters [1] and assist governments to better manage natural and infrastructural resources.

As an example of a spatial data centre and because of its affiliation to research, its international recognition and the advanced state of its data infrastructure we chose the Institut Cartogràfic de Catalunya (ICC) as a case study. The ICC creates maps and other geographically referenced data for the autonomic region of Catalonia (Spain) on a legal mandate. On local level other private and public institutions produce geodata related with Catalonia. The data are stored on local servers but the spatial data infrastructure of Catalonia (IDEC), a support department of the ICC, provides centralized viewing. The IDEC harvests metadata of map layers and data sets from the local servers to build its catalogue. The ICC is implementing the European INSPIRE directive which promotes a European spatial data infrastructure and explains how the data sets are to be described for full interoperability.

Through law 16/2005 about the geographic information and the ICC [4] the institution is obliged to preserve al cartographic material for future generations. This article presents the results of the bibliographic review on spatial data preservation which will serve as the knowledge base for the development of the preservation strategy for the ICC. These preservation strategies must be adapted to digital geodata, consider the whole data life cycle and take into account the decentralized production environment.

II. PARTICULARITY OF SPATIAL DATA

There are basically two types of spatial data: image and textual. The image can be vector graphics (maps and thematic layers) or raster graphic (remote sensed data as photography or digitized images). Vector graphics translates georeferenced data into points, lines, symbols and shapes [8]. Textual data often take the form of spread sheets. Data are organized in data sets, which might consist of homogeneous data representing one quality of a georeferenced feature as the thematic data layers in a geographic information system (GIS), or a continuous surface expressed in a collection of a single file type, as remote sensed imagery. In this article we want to focus on map layers in vector graphics as the problem of raster images is addressed broadly by image preservation projects of other industries. Particular to thematic spatial data layers is that they must be interpreted in context with a reference layer [3]. Spatial data are stored in different stages: raw, corrected, processed or published. Thematic data are in their raw form often numeric and take vector graphic shapes in their published form. In a GIS usually different resolutions and stages of the same data cohabit.

Spatial data preservation inherits all problems of digital preservation but has the following particular challenges:

- Need of partnership because of decentralized production.
- Often complex data that requires special knowledge for interpretation.
- The existence of different versions of the same data.

Each of these points is approached by one or two preservation projects in geosciences in collaboration with libraries and archives that will be presented in the next chapter.

III. SPATIAL DATA PRESERVATION EXPERIENCES

We identified few long-term digital preservation projects worldwide specifically focusing in geodata. Out of those, we chose four we considered best documented to compare their solutions.

A. National Geospatial Digital Archives, NGDA (USA)

The NGDA was a joint effort by the University of Stanford and the University of California Santa Barbara between 2004

and 2009. The two universities developed different technical and administrative solutions on a learning by doing bases. The project could build on the experience of the Alexandria Digital Library which disseminates georeferenced material through a distributed system. The collected experiences focused on legal solutions for producer-archive and archive-archive partnerships, shared collection development policies and format registry.

B. Geospatial Multistate Archive and Preservation Partnership, GeoMAPP (USA)

GeoMAPP started in 2007 and just concluded in December 2011. It was the natural continuation of a spatial data archival project in North Carolina form 2004 to 2007. As part of the National Digital Information Infrastructure and Preservation Program (NDIIPP) it investigated several digital preservation issues, including business planning, data inventory and metadata, appraisal and access [2]. GeoMAPP published a best practice for spatial archival processes and other outcome on its website[1].

C. VanMap (Canada)

The VanMap, a GIS system not meant to preserve even actualizations, became a case study in the context of the International Research on Permanent Authentic Records in Electronic Systems (InterPARES)[2] project in 2004. The first step of VanMap was to define whether or not an interactive map can become a record. The project especially focuses on archiving map layers in a way to maintain their usability and concentrates on automation of ingest and other processes. In collaboration with the San Diego Supercomputer Centre a preservation strategy was elaborated based on Grid-technology.

D. Swiss Federal Archives and swisstopo (Switzerland)

The Swiss Federal Archives (SFA) and the Federal Office of Topography (swisstopo) as the federal spatial data centre directed a study involving different spatial data stakeholders on government and local level. The existing data storage model of the MeteoSwiss "Data Warehouse" was analyzed and found not suitable for long term archiving of geodata. The study considered appraisal, selection criteria and snapshot frequency for geodata. During the project, a prototype for data transfer to the archive was developed taking into account metadata standards. Finally, the report proposes suitable file formats for long term archiving. The study finds continuation in the project Elipse[3] which is now concretizing the recommendations.

IV. SPATIAL DATA PRESERVATION CHALLENGES

All four above mentioned projects needed to find technical and legal answers to the three preservation challenges presented in chapter two. The following section is combining the solutions to each problem in order to paint a picture of the alternatives.

A. Decentralized production

All the archival projects are partnerships and insist in their benefit. Cultural institutions specialized in preserving information learn from data creators about the handling of spatial data and the producers learn about preservation needs. All studied cases require transfer agreements between institutions because the desired data are not produced in house. If public archives are the preserving stakeholder, the law fixes transfer, regulates access and eventually appraisal and retention frequency. Where such legal regulation is missing the transfer contract should include these aspects and define data ownership, authorized uses of the data and the responsibility and warranty of the archive towards the provider [13]. Additional guidelines on technical requirements and data quality can be added. A license model for institutions which deposit data to a library has been elaborated by the NGDA project. Partnership is also recommended between archives to complement collection policies and permit geographically remote storage of backup copies by partners.

B. Complexity of data.

There are two types of complexity: difficulty for human understanding and technical challenges for computer processing. For the purpose of this article we observe only complexity for human interpretation. Spatial data is especially sensitive to loss in human interpretability as it needs often special knowledge to understand it. Interpretability is the power of data to explain itself to a future user, once it has been taken out of its original context. Therefore, important archival work lies in description of data to provide context. The description is called metadata, data about data, and must be archived together with the vector graphics. The importance of context metadata for human understandability should not be underestimated as undocumented data sets must be considered useless [11]. Though, many different spatial data description standards are in use. In Europe the description standard "INSPIRE" [8] came into effect in 2008 and is mandatory for member states. By now, tools for creating INSPIRE-compliant metadata are on the market. Nevertheless, metadata creation is very time consuming, and accounts, according to the NGDA experience, for the biggest part of the archival process. As for the "GeoMAPP best practice" enhancing descriptive metadata at ingest is optional [10]. Nevertheless, the best practice prescribes that at least preservation metadata has to be added in order to maintain machine interpretability.

C. Managing versions of the same data

Most geosciences institutions use GIS to assist in decision making, data management and visualization. GIS generally overwrite data when it is actualized, or allow only poor recovery options for previous versions [2, 5, 9]. From the moment, an institution decides to track change over time, which is possible with newer GIS, it creates versions of the same data. An archived version is called "snapshot" or capture and represents the state of the data at a certain time. Captures can be planned on regular bases or be the effect of singular events such as new constructions or natural disaster. Not all versions must be captured for the archive. Erwin and Dingwell agree that the decision about the frequency must be taken for each data type individually.

Thematic geodata present a specific challenge in versioning: the thematic layers depend on reference data (underlying map) that can be actualized with different

[1] http://www.geomapp.net/
[2] http://www.interpares.org/
[3] http://www.bar.admin.ch/themen/00876/00939/index.html?lang=en

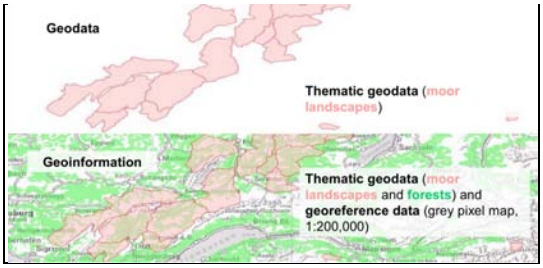

Figure 1. Tematic data layer without and with its reference layer. Reproduced by permission of swisstopo (BA12016)

frequencies. If the data do not have the correct base layer, it may cause wrong interpretation. The importance of the base layer is illustrated in 1. The Swiss geodata project [3] has analyzed three scenarios for the specific problem of thematic spatial data:

- Reference data are captured after each change to assure that the thematic data has always its corresponding reference. This might result in capturing reference data that is not needed.
- Every time a thematic data set is captured the corresponding reference data is archived as well. This might result in redundancy of reference data.
- Thematic and reference data are captured at a fixed frequency. This might result in thematic data that miss the correct reference data.

As all of the three variations have their disadvantages, Swiss preliminary study recommends using a combination. The frequency of archiving is important in regards to legal obligations where an archived record must reproduce an exact view of the data, as it was seen at a certain time. Important in this case is to maintain the link between the two layers so that a user can visualize them together.

V. Discussion

The ICC should first determine the organizational structure of its future archive. It has basically two options: 1) a centralized archive with the local governmental producers depositing the data at the ICC, or 2) maintaining the decentralization, unifying only the visualization by the user as it is the case with the current IDEC. In the first case the ICC could adapt and use the license model elaborated by the NGDA. The law gives the ICC the necessary freedom to place contracts and enter into partnerships. The contribution of local producers in Catalonia would be voluntary as they do not have legal obligation to deliver the data. In the second case the ICC would have no control on the preservation of the data stored at the local institutions. As their commitment to preservation might be varying the ICC could not guarantee the preservation of their data.

In either case the ICC has to find ways to guarantee interpretability of archived data sets. It should promote the use of the INSPIRE standard among the smaller ICC partners that did not implement it yet. The archive will not have the legal power to impose ingest requirements for datasets of suppliers except if it is agreed on in the transfer contract. Therefore, its role as mediator and advocator for the use of standards is crucial for the future understandability of the data. Bearing in mind that the archive maintains a catalogue with the harvested metadata, and meanwhile the standards are not completely implemented, the ICC has two options regarding metadata quality: it can decide to enhance and correct metadata itself, as do the two GeoMAPP partners Utah and North Carolina at ingest [9] or take over the metadata as transferred. The later is the case right now at the IDEC catalogue. If the ICC decides to trust in data suppliers and do nothing to control data quality it risks losing interpretability.

The third preservation challenge, versioning, is related to another function of the ICC. The Institute maintains a registry that certifies official map data. Any service that needs reference maps for its topic layer is obliged to use the official maps from the registry. Only if no official data of appropriate resolution is available the service can create its own cartography and apply for registry. Every institution that creates official reference data must supply a copy of the data to the ICC. The official map layers in Catalonia have legal value so that every registration of new data and removal of older versions are controlled by an administrative process. During the time the map has legal value al changes in this data must be recorded. By this means, archived topic layers of any institution will always have the correct reference data archived and its authenticity can be verified if the topic cartography was based on official data. After a certain period determined by law map layers lose their role as evidence. At this point the ICC must establish the capture frequency for long term archiving. If certain cartography does not have legal value it can be archived at larger intervals.

A centralized archive would allow the ICC to control the adequacy of the captured versions of reference and topic layers. Moreover the archive could standardize the files to allow combined use and ease interoperability. This option would take advantage of the existing technical expertise and infrastructure in the ICC offering it to other spatial data creators that might not have the personal and financial means to implement a trustworthy preservation system on their own. This way the ICC would enlarge its responsibility and influence on local data producers but also increases its work load and expenses. Finally, there might be a risk the ICC archive becomes a mere deposit of data other institutions do not want anymore.

The decentralized archival solution would leave more responsibility to the local producers. There would be more redundancy in data storage and probably more variety in software and platform solutions. Variety is favorable because it makes technical obsolesce more difficult. Also redundancy of data in geographically remote storage location is important in case of data loss. Though, in a decentralized system only part of the archived data have copies on several support media, in different file formats and on remote location.

In the actual context the metadata description lies in the responsibility of local creators. This way the ICC saves human resources and leaves specific topical knowhow to the data

authors while concentrating on the technical knowhow and infrastructure. Forcing little spatial data producers to create their own preservation system would not be bearable. Although, for a correct decision about centralized or decentralized archiving the administrative and legal consequences of both solutions must be subject of further research.

ACKNOWLEDGMENT

A.E.L. thanks Jordi Guimet, manager of the IDEC Support Centre, for his information and the Swiss project team of the SFA and swisstopo for the rights to publish the figure.

The reflections made in this paper are those of the authors and do not represent the official position of the ICC.

REFERENCES

[1] Beruti, V., Conway, E., Forcada, M. E., Giaretta, D., & Albani, M. (2010). ESA plans – a pathfinder for long term data preservation [online]. 7th International Conference on Preservation of Digital Objects (iPres 2010) September 19 - 24, 2010. Vienna. http://epubs.stfc.ac.uk/bitstream/6403/beruti-76.pdf. [Accessed 26/01/2012]

[2] Bethune, A., Lazorchak, B., & Nagy, Z. (2009). GeoMAPP: A geospatial multistate archive and preservation partnership. Journal of Map & Geography Libraries, 6(1), pp. 45-56. doi:10.1080/15420350903432630

[3] Bos, M., Gollin, H., Gerber, U., Leuthold, J., & Meyer, U. (2010). Archiving of geodata [online]: a joint preliminary study by swisstopo and the Swiss Federal Archives. http://www.swisstopo.admin.ch/internet/swisstopo/en/home/topics/geodata/geoarchive.parsysrelated1.59693.downloadList.93958.DownloadFile.tmp/preliminarystudyarchivingofgeodata.pdf. [Accessed 26/01/2012]

[4] Catalonia. Llei 16/2005, de 27 de desembre, de la información geográfica i de l'Institut Cartogràfic de Catalunya. DOGC no. 4543, 03/01/2006, p. 64.

[5] Dingwall, G., Marciano, R., Moore, R., & Peters McLellan, E. (2005). From data to records [online]: preserving the geographic information system of the City of Vancouver. Archivaria, 64(Fall 2007), pp. 181-198. http://journals.sfu.ca/archivar/index.php/archivaria/article/view/13157/14408. [Accessed 26/01/2012]

[6] Erwin, T., & Sweetkind-Singer, J. (2009). The National Geospatial Digital Archive: a collaborative project to archive geospatial data. Journal of Map & Geography Libraries, 6(1), pp. 6-25. doi:10.1080/15420350903432440

[7] Erwin, T., Sweetkind-singer, J., & Larsgaard, M. L. (2009). The National Geospatial Digital Archives—collection development: lessons learned. Library Trends, 57(3), pp. 490-515. doi:10.1353/lib.0.0049

[8] European Commission. (2007). Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE) [online]. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2007:108:0001:0014:EN:PDF. [Accessed 31/01/2012]

[9] Geospatial Multistate Archive and Preservation Partnership (GeoMAPP). (2010). GeoMAPP interim report: 2007-2009 [online]. http://www.geomapp.net/docs/GeoMAPP_InterimReport_Final.pdf. [Accessed 26/01/2012]

[10] Geospatial Multistate Archive and Preservation Partnership (GeoMAPP). (2011). Best practices for archival processing for geospatial datasets [online]. http://www.geomapp.net/docs/GIS_Archival_Processing_Process_v1.0_final_20111102.pdf. [Accessed 26/01/2012]

[11] Imfeld, S., Haller, R. Pitfalls in preserving geoinformation – lessons from the Swiss National Park. Preservation in digital cartography. Lecture notes in geoinformation and cartography, pp. 147-160. Doi: 10.1007/978-3-642-12733-5_7

[12] Janée, G. (2008). Preserving Geospatial Data [online]: The National Geospatial Digital Archive's approach. Archiving 2008: final program and proceedings, pp. 25-29. http://www.ngda.org/docs/Pub_Janee_Arch09_09.pdf. [Accessed 26/01/2012]

[13] Sweetkind, J.; Larsgaard, M. L.; Erwin, T. (2006). Digital Preservation of Geospatial Data. Library Trends, 55(2), pp. 304-314. Doi: 10.1353/lib.2006.0065