Miguel Termens, Anita E. Locher

32. Digital preservation audit on spatial data: a practical experience

Dept. of Library and Information Science, University of Barcelona
termens@ub.edu, *alocher@ub.edu*

Abstract: Digital spatial data are more vulnerable than traditional cartography due to their dependency on continuously evolving information technology, which makes them susceptible to obsolescence and accidental or malicious change.

The planning of digital preservation systems has to take into account the specific requirements of a service or institution. As such it is important that institutions can analyse internally or audit externally their ability to engage in and maintain digital preservation actions. Over the past few years, methodologies adapted to risk assessment and trustworthy digital repository audit have appeared, as is the case of Drambora, the Nestor criteria catalogue and TRAC (now ISO 16363:2012).

This article describes the experience of implementating an audit with TRAC in the "Institut Cartográfic de Catalunya (ICC)", in Barcelona, Spain. The ICC is currently exploring alternatives for long-term preservation of its data. Aiming at this goal, an audit conducted by external professionals was considered an excellent way to get an objective view of the ICCs current state.

Based on this experience and the execution of audits in other digital repositories (digital libraries...) the authors discuss the elements that should be considered for successfully carrying out audits in spatial data preservation systems.

Keywords: Digital preservation; Information Security; Trusted Digital Repositories; Repository Audit; Risk Assessment

32.1. Introduction

Digital formats have long been the preferred form of spatial data production, thanks to the fact that they can be easily handled and that, through interoperability, they make advanced geolocalisation services possible. However, digital data are more vulnerable than traditional cartographic data due to their dependency on information technology, in continuous evolution, which makes them susceptible to obsolescence and accidental or malicious change. Another

M. Jobst (ed.). Service Oriented Mapping 2012, 431-442

disadvantage for their long-term management is that digital assets are combined, shared and interacted with through specific online search and visualization services. In this sense the future preservation of spatial data infrastructures as IDEE', IDEC[2], the Geospatial Platform' and many others is a major challenge.

Spatial data preservation is the responsibility of all stakeholders, especially of producers, user communities and governments. The responsibility can change over time according to how archived spatial data are valued and used. There might be data of great immediate interest whose importance to their producers and original users may decline, but they subsequently can gain value for new users: for instance, those that study the temporal evolution of certain variables. These difficulties in clearly establishing responsibilities for data management may be the cause of stakeholders' current nonchalance regarding the future preservation of data.

At a technological level there is a better outlook as there is a generally accepted conceptual frame —ISO 14721:2003 OAIS— (ISO 2003, Consultative 2012) that serves for planning digital preservation systems. However, it must adapt to the specific requirements of a service or institution. Digital preservation systems are technically complex and need to be managed by specialized professionals. They depend as well on solid financial and organizational support. Realistically, it is clear that not all the institutions will be able to carry out preservation activities independently, but rather collaboration will often be necessary.

Whether digital preservation activities are the responsibility of a single institution, are performed by distributed network services, or are managed by a consortium or group of institutions, it is not only necessary to ensure the technological reliability of the activities, but the sustainability of funding and management mechanisms in the short and medium term must also be guaranteed

Ultimately it comes down to ensuring the availability of a secure and reliable archive or repository. This reliability must be checked over time to prove that preservation functions are being conducted properly. Auditing systems already tested in other service areas are most suitable for checking the reliability of a digital preservation system and therefore must be incorporated into the mechanisms of spatial data preservation.

Below we introduce various audit systems applicable to institutions that preserve digital data. Subsequently, we present an audit experience about the trustworthiness of a repository applied to the spatial data production centre, the "Institut Cartogràfic de Catalunya (ICC)", in Barcelona, Spain.

http://www.idee.es/

z http://www.geoportal-idec.cat/geoportal/cat/

http://www.geoplatform.gov/home/

Methodological conclusions can be drawn from the audit results which could subsequently be applied to other institutions in the same area.

32.2. Alternatives

The PDCA methodology (Plan-Do-Check-Act) has been successfully applied to quality management (ISO 9001:2008) and information security management systems (ISO 27001:2005) among others. A key feature of audit methodologies is that they can check the proper functioning of management systems.

Verification through audit can be external or internal. For an outside observer, passing the audit may seem the principal goal of the institution in conducting it, but from an internal point of view they are mechanisms that allow errors to be detected, provide a neutral view of system inefficiencies and, as a result, help in drafting improvement plans.

During the last decade digital preservation systems have been implemented in various fields, especially in libraries, archives and publishing. In some cases these are commercial or joint services that allow outsourcing of digital content preservation. These services take the form of repositories of different technological architectures — centralized, as a network or in the cloud - which contain large volumes of data and they must be able to ensure their trustworthiness to partners or customers (Ross and McHugh 2006, Dryden 2011).

Although preserving digital content responds first of all to a technological problem - the obsolescence of formats, hardware and software that can prevent future access to data — in reality its implementation is not only an IT issue; answers to legal, financial and administrative questions regarding the system must be sought as well,. For this reason the audit methodologies of computer systems, among which is the ISO 27001:2005, are not sufficient as they focus on just one aspect of the problem.

In this context, several methodologies adapted to risk assessment and trustworthy digital repository audit have appeared and are being put into practice. These include Nestor, Drambora and TRAC. In 2007 the NESTOR (Network of Expertise in Long-Term Storage of Digital Resources) working group under the Ministry of Education and Research of Germany developed the Nestor Catalogue of Criteria for Trusted Digital Repositories (Dobratz et al 2007; Catalogue 2008); the second version, published in 2009, consists of a structured list of instructions a digital repository must comply with in order to be considered secure. The principle criteria are

- adequacy —inexistence of absolute standards,
- measurability —all indicators have to be measurable,
- documentation —the objectives, specifications and implementation have to be documented,
- transparency —part of the documentation has to be public in order to inspire trust in stakeholders.

Currently, Nestor is the method of reference for auditing repositories for scientific publications and other content at German universities.

Drambora (Digital Repository Audit Method Based On Risk Assessment) arose from a research project funded under the framework of the European Union and led by the University of Glasgow's Humanities Advanced Technology and Information Institute (HATII) and the National Archives of The Netherlands (Quisbert 2008). It takes the form of a checklist that allows for the confirmation of the level of implementation of digital preservation policies in a repository. The scope of Drambora is narrower than that of Nestor, as evidenced by the fact that it is not intended for external audit and certification but focuses on internal audits.

TRAC (Trustworthy Repositories Audit & Certification: Criteria And Checklist) starts from studies initiated in 2005 by OCLC and the Center for Research Libraries of the United States (Hank et al 2007). TRAC, published in 2007 (OCLC 2007), is a complete method with a form similar to Nestor, developed as an external audit mechanism. It has been tested in some of the main preservation repositories in the United States —Portico, Hathitrust and Chronopolis- (Center for Research Libraries 2010, 2011, 2012). This experience led to a new version of TRAC under the name of TDR (Trustworthy Digital Repository Checklist), (Consultative 2011b) which was recently approved as ISO 16363:2012 standard (ISO 2012). It is expected that it will soon be possible to conduct certification audits under this standard.

The Geoarchiving Self Assessment tool (Geospatial 2010) is a checklist developed by the GeoMAPP[4] project partners and addressed to geoscience institutions that store their own data in order to enable them to objectively evaluate their archives' potential to preserve geospatial data or test their repositories current archiving practices. The criteria are classified in three categories from basic to advanced. Each of these levels covers the following areas: Plan sponsorship and project governance, current programs, human resource requirements, data requirements, and technological requirements. The questions about these subjects are very detailed, but are not accompanied by indications on how to answer the questions, as in TRAC for example. As it is a recent product, there are no reports or experiences concerning the Geoarchiving Self Assessment tool available yet.

32.3. Problem

The ICC is the main government entity producing location and time referenced data for the autonomic region of Catalonia (Spain) by legal mandate. Their reference and thematic map production includes different scales between 1:5,000 and 1:25,000. This 30-year old institution switched to full digital production in 2005. Commercial interest and a legal mission to preserve the map heritage of Catalonia for future generations led to its policy of never deleting any data.

[4] GeoMAPP is an abbreviation for Geospatial Multistate Archive and Preservation Partnership (http://www.geomapp.net)

Data are managed by the internal data centre which stores long-term data on storage tape and makes system back-ups. At the data centre born digital maps are stored with the digitized versions of older paper maps. The sheer amount of archived data made the ICC reflect on the necessary preservation actions and the creation of technological instruments and internal procedures to make data preservation feasible on the long term (Anguita et al 2012).

The Library and Information Science Department of the University of Barcelona took part in these reflections. The collaboration consisted of conducting an audit of the trustworthiness of the ICC as a digital repository. The ICC expects that this method will help identify its level of preparation for properly conducting digital preservation tasks on spatial data.

We report this experience, as it is one of the first audits made with TRAC specifically in a geoscience institution and in order to highlight the differences that appear by auditing an institution that stores its own data compared to cultural institutions that store data from third parties.

32.4. Methodology

The team of auditors was composed of a doctoral student and a professor from the University of Barcelona, both specialized in digital preservation. The relationship between auditor and auditee has been kept as independent as possible in two senses: independent in mind and independent in appearance (Porter et al 2003). Independence in mind is achieved by the fact that the two bodies are not dependent on the same political structure and therefore are not pressured by each other. Likewise there is no commercial interest on the part of the auditors which could influence their opinion nor was there any legal pressure on the auditee to conduct this audit. In order to create trust in the auditor and maintain a high level of independence, the audit has been voluntary for the ICC and it has received a declaration of confidentiality signed by the auditors stating the risks involved, as indicated in TRAC.

On the one hand, it is clear that our audit team did not meet the requirements for certifying bodies since we haven't had the sufficient experience in audit and risk assessment as required in 7.2.2.3 paragraph c. Also, despite the preparatory meetings and studies of the context, we lacked a full understanding of the client organization before starting the audit. Nevertheless this second disadvantage was overcome by clarifying the technical structure and procedures during the interviews. We are not a certified auditing company as described in ISO 17021 and Requirements for Bodies Providing Audit and Certification of Candidate Trustworthy Digital Repositories: recommended practice CCSDS 652.1-M-1. On the other hand, preconditions on the auditee side were promising as we had management support and a dedicated ICC staff member who is an expert on the repository. These are two criteria that Walls (2011) considers essential for a successful audit.

The ICC has a department responsible for quality control and maintains the ISO 9001 standard for several of its products; therefore it is accustomed to auditing processes. Nevertheless, this was the first time TRAC has been applied to its digital collection since the creation of the repository.

In a preparatory phase we translated the TRAC criteria to Spanish. In the first meeting with a member of the ICC, we sealed the collaboration and in a second meeting, we tried to clear up all the doubts we had about the technical and organizational structure of the institution. A third meeting was held to explain to concerned ICC members the three-fold structure of TRAC and the methodology we would use and also to get the approval of the managing directors. It also provided the occasion to get to know the people we would interview during the audit.

Reading the requirements for auditing bodies (Consultative 2011a) was helpful in allowing us to better understand attitudes and basic knowledge requirements. In order to complement the information received in the meetings, we studied the law (Generalitat 2005) and regulations in order to have an adequate understanding of the regulatory requirements applicable to the ICC. With this acquired knowledge and with the help of the institutional web page, we finally adapted the TRAC questions to the vocabulary and structure of the ICC.

As the certification of the ICC was not the goal, the audit procedure was simplified to the point that the auditors did not search for proof, but merely relied on the accuracy of the answers provided by the auditee. On this point our method differs from usual external audit practice that requires availability and enquiring of documentation in order to confirm the obtained information. Likewise, many questions that could lead to a "yes or no" answer were changed in order to get more information about the "how".

The TRAC checklist was answered by the director of the IT department in two personal interviews with the auditor team. The director of the IT department is especially knowledgeable about digital preservation issues as it is he who leads the ICC's efforts on this topic. Questions that could not be answered by the director of IT were forwarded by him to appropriate staff. The answers collected in this way were transmitted to the auditor in the second interview. Following the interview the auditee clarified the doubts of the audit team by e-mail.

32.5. Results

Although we cannot give specific information about the audit results for reasons of confidentiality, we will briefly highlight situations that have occurred at other audits of science institutions that store their own data.

During the preparatory meetings with the ICC we were pleased to see that we shared a great part of the technical vocabulary. Despite its relatively short history (30 years), the ICC has already gone through several digital preservation challenges such as the recovery of files from a tape for which readers were no

longer available and the temporary loss of a great amount of data. These experiences provided expertise for the Institute in digital data handling and awareness of preservation concerns. Furthermore growing costs and amounts of data resulting from the ICC's determination to not discard any data, made it very receptive to the audit request.

The willingness of the geoscience community to share information has also had a positive influence on the understanding between auditor and auditee. Indeed the Infrastructure for Spatial Data in Catalonia (IDEC), which is part of the ICC, promotes data sharing between geodata producers. Metadata harvesting and international standards are their daily business as well as being crucial for preservation.

Nevertheless, TRAC uses specific concepts which require understanding of the OAIS model (ISO 14721:2003), which led us to the decision to not supply the TRAC questions to the ICC prior to the interviews. There were two reasons for this: firstly, we think it saves the clients time to answer oral question because any doubts can be clarified right away. As the ICCs participation was voluntary but not a high priority, we wanted to make it as easy for them as possible. Second, because it was the ICC's first contact with the standard and its vocabulary, we wanted to avoid misunderstandings that could occur if the clients had to answer the questions by themselves. The higher investment on the auditor side was meant that sufficient time had to be reserved for the interviews.

Below we detail points of the TRAC checklist where potential conflicts of definition arose because the ICC stores in-house production. When auditing this kind of repository, the concepts need special attention and have to be defined together with the auditee.

32.5.1. Mission statement

Obviously, preservation is not the primary mission of scientific institutions, resulting in mission statements that lack the necessary mention of digital preservation (TRAC criteria 3.1.1). Production, service and dissemination are probably the most important aspects for these institutions. However, the institution can have a department dedicated to digital preservation with its own mission, though this is not equally reliable, since preservation issues are thereby on a lower level in the consciousness of CEOs. Additional trust in the long-term commitment of the institution is built when the institution's members rely on the collection for doing longitudinal research, clients need to access their own stored data or, as is the case of the ICC, the responsibility for preservation is based on law.

32.5.2. Licensing

It can be tempting to dismiss answering the questions about licensing (TRAC criteria 3.5) when the production is all in-house. However there are almost always licensing concerns about the software, file formats or hardware.

32.5.3. Ingest

TRAC uses the vocabulary of the OAIS model because it expects a repository to conform to the OAIS structure. OAIS indicates that information submitted to an archive should be packaged in order to hold together intellectual entities. Every package contains one or more files: typically data and associated metadata. At the ICC, there is neither a formal submission, as the institution stores its own data, nor is there a packaging process, which is why we considered the single file as a package. The fact that the file itself is the storage unit, the unique file name can be understood as the package identifier. Because there are no packages, metadata has to be tied to the data in other ways, such as being embedded in the file or related to it by the unique identifiers. The former is more appropriate for provenance metadata that is automatically created at production; the latter might be more useful as context metadata (TRAC criteria 4.2.1.2).

According to OAIS, the archival unit is a similar package whose files have gone through preservation actions such as virus scans, verification of integrity by comparing checksums, normalization of the file format, quality control etc. The only preservation action taken at this point at the ICC is creation of a checksum. Calculation of a checksum does not alter the file. As there are no other preservation actions, the files themselves represent both the submission and the archival packages as they are identified and stored by the IT department. Therefore, the submission and archival formats are identical, which made some points of the checklist redundant, such as the identification of the submission package (TRAC criteria 4.1.1) and the archival package (TRAC criteria 4.2.4.1).

32.5.4. File exclusion

An internal IT department cannot refuse to store data, which is why it has no protocol for action in case of file exclusion (TRAC criteria 4.2.3.1). However this would change if different storage environments were created for long-term storage and short- or medium-term storage, leading to the establishment of criteria to determine how to separate the types of data according to life cycle requirements. Separation into two or more repositories might be useful for primary science data production (long-term storage) and administrative data (documents and media that have been used in order to produce the primary data).

32.5.5. *Adequacy of methods*

The understanding we gained about the current storage practice, the challenges a geodata producer faces, and the ICC's organizational structure was fully satisfactory for the audit team. The fact that in the beginning, we expected three repositories for items of different provenance resulted in some inappropriate answers. Nevertheless, the extensive explanations of all the functions and processes of the repository and IT department that were the result of subsequent questions, gave us enough information to clarify earlier misunderstandings. Through this audit we gained an overview of the state of preservation actions at the institution. The three parts of the TRAC checklist guaranteed

that we not only concentrated on the technological aspect of preservation but could also see how far up digital preservation issues reached in the consciousness of the governing levels of the institution. This will help us estimate the effort needed by the institution to create a trusted preservation system and compare it to the probability and effort other local geodata producers will have in setting up such a system.

The fact that we worked with the magenta book (a version of TRAC previous to the ISO publication) did not influence the audit. The ISO 16363 standard does not contain changes compared to the version we used.

In a case where TRAC is applied for the first time, it can be useful to complete the audit with the DRAMBORA tool in order to get deeper insight into all related risks as required in criteria 5.1.1 of the checklist. Other standards (for example ISO 9001 for quality management, ISO 15489 for records management and ISO 17799 for data security) are similar to TRAC and in some points are even overlapping (TRAC criteria 5.2.2, 5.2.3 and 5.2.4). However, if it is the production process that is certified, this has little significance over the trustworthiness of the repository.

32.6. Discussion

To date, TRAC has been used to audit preservation systems based on a centralized repository under the responsibility of a single institution — with the exception of Altman and Crabtree (2011) - and was the case in our experience auditing the ICC. We don't know if TRAC would also fit distributed architectures based on cloud services. Until now, this type of preservation systems has been found only in experimental phases at cultural heritage institutions, such as libraries and archives, which is why their reliability in the long term has not been proved yet. However, distributed architectures are usual for spatial data services that integrate layers coming from different producers with unified search and retrieval interfaces.

The ICC is responsible for coordinating the Spatial Data Infrastructure of Catalonia (IDEC), which has a decentralized nature. Therefore, preservation planning for the ICC's data will have to take into account the relation with the preservation of all data accessible through the !DEC. In a data sharing structure provenance and licenses would take on greater importance. Individual data providers would have to agree to license terms giving the ICC sufficient control over the data for it to perform preservation actions which could alter the data (criteria 3.5). In an environment consisting of data producers of different sizes and levels of technological capability, audit methods can be used as tools to discover weaknesses in the system and, as a result, establish the most appropriate policies. In this context the experience of Steinhart and Dietrich (2009) who used TRAC for the design of a staging repository is interesting, since the IDEC could play this role in the future: it could offer a platform where the harvested data would be enriched with necessary metadata and prepared for ingest in other repositories or distributed as is already the case. A preservation environment for geodata in Catalonia could use TRAC to design such new functions of an IDEO staging repository and to determine the trustworthiness of the repositories receiving the final data. The greater the number of repositories reaching a certain degree of trustworthiness, the greater the number of reliable options that producers will have for depositing their data. Auditing of individual repositories at the end of the preservation chain can be speeded up when functions of the repository software are related to TRAC questions. Efforts in this direction have been undertaken by Kaczmarek et al (2006). Similar research should be done for geographic information systems so that the results could help in evaluating trust in smaller geospatial repositories that are part of the preservation chain.

TRAC (now ISO 16363) is applicable to all kinds of digital data and not limited to geodata. We recommend to other institutions that have to store their own data in a repository to use this standard's checklist to monitor their progress in reaching trustworthiness. They will encounter similar situations as described in our results that will force them to adapt the terms used in the standard to their own context.

32.7. Acknowledgements

The authors thank the staff of the ICC for their collaboration in the audit.

References

Altman, M, Crabtree J. (2011) Using the SafeArchive System: TRAC-Based Auditing of LOCKSS. In: Archiving 2011 final program and proceedings. Salt Lake City, USA 16 - 19 May 2011. Springfield: Society for Imaging Science and Technology.

Anguita, S., Montaner, C., Oiler, J. & Roset, R. (2012) Digital preservation at the Institut Cartogràfic de Catalunya. [CD-ROM]. In: 7th International Workshop Digital Approaches to Cartographic Heritage. Barcelona, Spain 19-20 April 2012. Barcelona: Institut Cartográfic de Catalunya.

NESTOR (2008) Catalogue of criteria for trusted digital repositories: version 2. [pdf] Frankfurt: Nestor Working Group Trusted Repositories - Certification. Available at: <http://files.d-nb.de/nestor/materialien/nestor_mat_08_eng.pdf> [Accessed 26 July 2012].

Center for Research Libraries (2010) Report on Portico Audit Findings. [pdf] Available at: http://www.crl.edu/sites/default/files/attachments/pages/CRL Report on Portico Audit 2010.pdf [Accessed 26 July 2012].

Center for Research Libraries (2011) Certification Report on the HathiTrust Digital Repository. [pdf] Available at: <http://www.crl.edu/sites/default/files/attachments/pages/CRLHathiTrust 2011.pdf> [Accessed 26 July 2012].

Center for Research Libraries (2012) Certification Report on Chronopolis. [pdf] Available at: <http://www.crl.edu/sites/default/files/attachments/pages/Chron_Report_2012_final_0.pdf> [Accessed 26 July 2012].

Consultative Committee for Space Data Systems (CCSDS) (2012) Reference Model for an Open Archival Information System (OASIS). Recommended Practice. Magenta Book. [pdf] Washington DC: CCSDS. Available at: <http://public.ccsds.org/publications/archive/650x0m2.pdf> [Accessed 26 July 2012].

Consultative Committee for Space Data Systems (CCSDS) (2011a) Requirements for Bodies Providing Audit and Certification of Candidate Trustworthy Digital Repositories: recomended practice. [pdf] Washington DC: CCSDS. Available at: <http://public.ccsds.org/publications/archive/652x1m1.pdf> [Accessed 26 July 2012].

Consultative Committee for Space Data Systems (CCSDS) (2011b) Audit and certification of trustworthy digital repositories: recommended practice CCSDS 652.0-M-1. Practice. [pdf] Washington DC: CCSDS. Available at: <http://public.ccsds.org/publications/archive/652x0m1.pdf>

Dobratz, S., and Schoger, A. (2007). Trustworthy Digital Long-Term Repositories: The Nestor Approach in the Context of International Developments. Lecture Notes in Computer Science, [e-journal] 4675, 210-222 Available through: <http://link.springer.com/chapter/10.1007/978-3-540-74851-9_18> [Accessed 2 October 2012].

Dryden, J. (2011) Measuring Trust: Standards for Trusted Digital Repositories. Journal of Archival Organization, [e-journal] 9(2), 127-130. Available through: Education Resources Information Center <http://www.eric.ed.gov/> [Accessed 26 July 2012].

International Standards Office (2003) ISO 14721:2003. Space data and information transfer systems - Open archival information system - Reference model. Geneva: ISO

International Standards Office (2012) ISO 16363:2012. Space data and information transfer systems - Audit and certification of trustworthy digital repositories. Geneva: ISO

Generalitat de Catalunya (2005) Llei 16/2005, de 27 de desembre, de la informació geográfica i de l'Institut Cartográfic de Catalunya, Diari oficial de la Generalitat de Catalunya, [online] 4543. Available at:

<https://www.gencat.cat/eadop/imatges/4543/05361027.pdf> [Accessed 2 October 2012].

Geospatial Multistate Archive and Preservation Partnership (2010) Geoarchiving Self Assessment [spread sheet] Available at:

<http://www.geomapp.net/docs/GeoMAPP_GeoArchiving_SelfAssessment_20100914.xls> [Accessed 26 July 2012].

Hank, C., Tibbo, H. R., and Barnes, H. (2007) Building from trust: using the RLG / NARA Audit Checklist for institutional repository planning and deployment. In: Archiving 2007 Conference Proceedings. Arlington, USA 21-24 May 2007. Springfield: Society for Imaging Science and Technology.

Kaczmarek, J., Hswe, P., Eke, J. and Habing, T. (2006) Using the audit checklist for the certification of a trusted digital repository as a framework for evaluating repository software applications: A progress report. D-Lib Magazine [online] 12(12). Available at: <http://www.dlib.org/dlib/december06/kaczmarek/12kaczmarek.html> [Accessed 30 July 2012].

OCLC (2007) Trustworthy Repositories Audit & Certification: Criteria and Checklist Version 1.0. [pdf] Available at: <http://www.crl.edu/content.asp?l1=13&l2=58&l3=162&l4=91> [Accessed 26 July 2012].