Tools for a long-term preservation of digital documents in trustworthy repositories

Andrea Fojtu andrea.fojtu@ruk.cuni.cz
Eliska Pavlaskova eliska.pavlaskova@ruk.cuni.cz
Jan Hutar jan.hutar@nkp.cz

Charles University of Prague. Czech Republic

National Library of the Czech Republic. Czech Republic

Abstract: The paper presents the process of creating a long-term preservation plan. This should start with reviewing state-of-the-art of the repository, then continue with choosing the appropriate formats, methods, SW, HW, testing of chosen policies and auditing and certification. Institutions' efforts can be supported by a wide variety of tools, e.g. Platter, Plato, DRAMBORA, Nestor, TRAC. Most of them were tested and used at the National Library of the Czech Republic and Charles University in Prague.

Keywords: long-term preservation, OAIS reference model, preservation plan, mind-maps, Platter, Plato, DRAMBORA, Nestor, TRAC

Information is produced in a wide variety of formats. Born-digital and digitized data has become an integral part of most of the world's cultural heritage collections.

Increasing amount of digital objects makes librarians and other information specialists aware of a digital objects specific nature in the context of digital preservation needs. Preservation principles applicable on traditional documents are diametrically different from the principles for digital objects. Current state of research in this field suggests necessity of preliminary and continuous actions more than just a damage management.

Worldwide digital preservation community has focused its attention on a variety of complex problems regarding preservation of both born-digital and digitized objects. During the last few years intense research projects and useful tools have been conducted and developed. Activities of the community also concentrate on maintaining enduring authenticity, understandability, renderability, viability, integrity, identity and availability of at least essential characteristics of digital documents.

All of the digital preservation working groups adopted OAIS (Open Archival Information System) Reference Model as a source of terminology and general theoretical framework for digital preservation. The Model defines digital archive as an organization responsible for digital preservation ("digital repository" can be used for description of mostly identical concept) and the term "designated community" for "identified group of potential consumers who should be able to understand a particular set of information." [3]. Three information packages were also identified in the process of archiving and preserving of digital material – SIP, AIP and DIP (Submission Information Package, Archival Information Package and Dissemination Information Package). The model also describes five functional entities – Ingest, Archival Storage, Data Management, Administration and Access.

Producers who are part of the information system environment create the Submission Information Packet. Typically, SIP includes information about content of the object (e.g. set of files), but significant part of preservation information (e.g. technical metadata) need to be added. Preservation information includes preservation information necessary for object archiving. AIP is stored in information system and repository is responsible for its preservation. Dissemination Information Package is represented to consumers (users) and its form depends on user-needs and presented medium.

At the beginning of defining an internal/institutional's long-term preservation of digital documents it is necessary to start with reviewing state-of-the-art of the repository (see A. An initial point), then continue with choosing the appropriate formats, methods, SW, HW (see B. Choosing), testing of chosen policies (see C.

Testing) and auditing and certification (see D. Certification and audit).

A. AN INITIAL POINT

State-of-the-art of the institutions' repository could be quite easily captured with the mind-map tools like FreeMind or Bubbl.us.

FreeMind [6]

Owing to the fact that the software is written in Java, it is available for many operating systems (MS Windows, Mac OS X, Linux, eComStation platforms). Even though it exists as a client application only (without the online counterpart), its indisputable advantage is integration into the planning tool PLATO (see the text below). However, if the organization is primarily looking for an application for brainstorming sessions and ideas sharing, Bubbl.us is the best choice.

Bubbl.us [1]

Bubbl.us presents an easy-to-use, free and online tool for creating mind maps. Mind maps can be shared, posted to a web site or blog, sent via e-mail, printed or saved as an image.

Thanks to these tools we understood the problems in their very context and could continue in defining feasible formats of documents ingested into the repositories (giving up the idea of storing and preserving everything), since this should be an integral part of the institution's preservation strategy.

B. CHOOSING

When choosing the right strategy (for a long-term preservation) we had to reassess the acquisition of certain file formats. According to the Florida Digital Archive recommendation of data formats for preservation purposes [5], proprietary formats in general are not acceptable.

Platter – Planning Tool for Trusted Electronic Repositories [7]

Long-term preservation of digital documents issues should start at the very first stages of a repository planning process.

PLATTER in itself is not an audit or certification tool, but is rather designed to complement existing audit and certification tools by providing a „cookbook", which will allow new repositories to incorporate the goal of planning the long-term preservation issues from an early stage. This shows the scope of questions and risks, which may arise and need to be addressed in order to comply with a trusted repository (sometimes mistakenly denoted as trustedness of repository) [4].

And the repositories are trusted if they meet predefined criteria. Fulfilling the functions or (if you like) meeting the criteria needs to be declarable. That means having the trusted repository is considerably subject to certification and audit (see part D. Certification and audit). PLATTER has been designed to support both checklist and risk-analysis based approaches to audit.

The biggest advantage of PLATTER is identification of strong and weak points of our present (or future) repository. Unrealistic expectations, impossibility of application of all requirements to any repository and in some cases too simplistic and general recommendations should not prevent us from looking for the finest solutions (see C. Testing and D. Certification and audit).

C. TESTING

In next to the last step of creating the plan for the long-term preservation of digital documents we focused our attention to testing (of chosen formats, solutions etc.). For this we used the tool called Plato.

Plato – The Planets Preservation Planning Tool [9]

Plato supports decision making in the field of preservation planning. The tool is based on OAIS functional entity Preservation Planning. It is a software tool (licensed under the CC-GNU LGPL) for evaluation of potential preservation solutions and strategies.

Whole process is divided into three stages and eleven steps described by workflow of decision making:

1. ***Define requirements*** – this consists of defining basis, where basic information about collection and its context is inserted. This step is similar to the first phase of DRAMBORA audit. With the second step, records need to be chosen. It requires a sample set of records for later use. Identify Requirements, as a third step, deals with repository objectives definition and description. It is recommended to use a tree structure for objectives categorization. Top-level categories usually are file, record as well as process characteristics and costs.

2. ***Evaluate alternatives*** – it includes definition of alternatives (all possible preservation solutions). After the decision is made, it is time to focus on feasibility of chosen alternatives. Feasible solutions represent the evaluating process and become a subject of experiment phases.

3. ***Consider results*** - management is provided with experiment output included measurements taken in the experiments. Measurements can be transformed into a uniform scale. In case when not all of the objectives are equally important, importance factors can be set. In the last phase relevant results need to be analysed.

The output of preserving planning process is "a concise, objective, and well-documented ranked list of the various alternative solutions for a given preservation task considering institution-specific requirements" [8]. The list can be used as a basis for optimal preservation solution decision (for a given collection).

D. CERTIFICATION AND AUDIT

The last phase before the final version of long-term preservation plan/strategy of digital documents should be a complete audit of previous policies, procedures and results. Most often with the help of NESTOR, TRAC or online tool DRAMBORA Interactive; choosing the most suitable one is subjected to human resources, material and last but not least financial possibilities.

NESTOR - Network of Expertise in Long-Term Storage and Long-Term availability of Digital Resources in Germany [2]

It consists of 14 criteria with detailed description and specific application. The trustworthiness evaluation and certification of a repository with the help of the NESTOR consists of 3 parts: 1. ***Organisation*** (administration, sustainability, institution structure and provision), 2. ***Digital object and technology management***, 3. ***Technical infrastructure*** and ***security***.

NESTOR's target groups include research institutes at universities, in industry and within specialist organisations. Further targets include libraries, museums and archives as users.

The reason we did not work with these criteria in more detail is their close relationship to legal and financial conditions and limitations in Germany.

TRAC - Trustworthy Repositories Audit & Certification : Criteria and Checklist [10]

This paid certification has already been conducted at several institutions (e.g. USA, Netherlands, New Zealand). Due to the rather high financial requirements, we preferred the „self-audit" by DRAMBORA.

DRAMBORA - Digital Repository Audit Method Based on Risk Assessment [11]

As its title shows, DRAMBORA is a method based on risk management. Its main purpose is to work as a tool for repository audit or self-audit and eventually as a planning tool prior a launch of repositories. It adopts a bottom-up approach for identification of potential threats. The tool was created in the beginning of 2007 as a cooperative project of Digital Curation Centre (DCC) and Digital Preservation Europe (DPE). Since January 2008, DRAMBORA is available as online tool called DRAMBORA Interactive.

The online tool guides auditors through stages of (self-) audit methodology. It is strongly recommended to conduct preliminary analysis of existing repository documentation not only in written form (the tool provides functions for managing documents significant for repository), but also in a form of interviews with responsible repository personal [4]. The audit has six phases; starting with organization mandate definition and organization context description and ending with two distinct outputs from assessment process - register of indentified risks and structured audit report. Risks, vulnerabilities are identified, measured and relations between them are recognized.

DRAMBORA was tested in several institutions including the National Library of the Czech Republic and the Charles University in Prague. The main benefit of the conducted audit was not in new information about problematic issues but structured and systematic overview of known problems.

E. CONCLUSIONS

It would be a mistake to presume that the long-term preservation of digital documents is easy to be solved by a metadata description and simple storage of date on (optical or solid) media. At present the issue of the long-term preservation is no longer only related to technologies. It includes organisation, administration, management, a qualified personnel and financial resources.

1 *Bubbl.us* [online]. [cit. 2009-10-15]. Available at: <http://www.bubbl.us/>.

2 *Catalogue of Criteria for Trusted Digital Repositories*. Version 1. Goettingen : Nestor Working Group, 2006. 48 s. Available at: <http://edoc.hu-berlin.de/series/nestor-materialien/8/PDF/8.pdf>.

3 Consultative Committee for Space Data Systems. *Reference model for an open archival information system (OAIS)* [online]. Washington (DC) : National Aeronautics and Space Administration, 2002 [cit. 2009-09-20]. Available at: <http://www.ndk.cz/dokumenty/650x0b1.pdf>.

4 DONELLY, M., INNOCENTI, P., MCHUGH, A., et al. *Drambora Interactive User Guide* [online]. Glasgow : Digital Preservation Europe, 2009 [cit. 2009-09-20]. Available at: < http://www.dcc.ac.uk/docs/tools/DRAMBORA_Interactive_Manual.pdf>.

5 *Florida Digital Archive* [online]. Gainesville, FL : Florida Digital Archive, c2003 [cit. 2009-05-09]. Recommended Data Formats for Preservation Purposes in the Florida Digital Archive. Available at: <http://www.fcla.edu/digitalArchive/pdfs/recFormats.pdf>.

6 *FreeMind - free mind mapping software* [online]. 2004 , last modified 14:04, 6 Sep 2009 [cit. 2009-10-15]. Available at: <http://freemind.sourceforge.net/wiki/index.php/Main_Page>.

7 *Repository Planning Checklist and Guidance* [online]. DigitalPreservationEurope, 2008 [cit. 2009-10-15]. Available at: <http://www.digitalpreservationeurope.eu/platter.pdf>.

8 STRODL, S., BECKER, C., NEUMEYER, R., RAUBER, A. *How to Choose a Digital Preservation Strategy: Evaluating a Preservation Planning Procedure* [online]. June 2007 [cit. 2009-04-05]. Available at: <http://ifs.tuwien.ac.at/~strodl/paper/FP060-strodl.pdf>.

9 *The Planets Preservation Planning Tool* [online]. [2007] [cit. 2009-10-15]. Available at: <http://www.ifs.tuwien.ac.at/dp/plato/intro.html>.

10 *Trustworthy Repositories Audit & Certification (TRAC): Criteria and Checklist*. Dublin, OH; Chicago, IL : OCLC, CRL, 2007. 94 s. Available at: <http://www.crl.edu/PDF/trac.pdf>.

11 *Welcome to DRAMBORA Interactive: Log in or Register to Use the Toolkit* [online]. 2008. Published 01-02-2008. [cit. 2009-09-20]. Available at: <http://www.repositoryaudit.eu>.