LIBER QUARTERLY 19 (1), April 2009 – ISSN: 1435-5205. P13–21
http://liber.library.uu.nl/
Igitur, Utrecht Publishing & Archiving Services

The Jigsaw Puzzle of Digital Preservation — an Overview

Barbara Sierman

Team Leader, Digital Preservation Research,
Koninklijke Bibliotheek, National Library of the Netherlands,
PO Box 90407, 2509 LK The Hague,
barbara.sierman@kb.nl

Abstract

Before the 22nd Annual Meeting of the Board of Directors of the Foundation CENL, Zagreb, September 24–27, 2008, the author presented a clear overview of the latest developments in digital preservation in a European context. She dealt with organisational aspects, the digital objects themselves, and the effects of international European collaboration. She calls on European organisations such as the Alliance for Permanent Access to sustain the results of temporary projects like PLANETS and thereby bring the pieces of the digital preservation puzzle together.

This paper is being published in preparation of the workshop on Curating Research: e-Merging New Roles and Responsibilities in the European Landscape, which is being co-organised by LIBER on 17 April 2009 at the Koninklijke Bibliotheek in The Hague.

Key Words: digital preservation; CENL; libraries; European projects

Introduction

Digital preservation is like a jigsaw puzzle: a nice box with thousands of pieces in it and a beautiful picture on the outside, which you can see if all the pieces of the puzzle are put together in the right way, often after a tremendous lot of effort and perseverance. The digital preservation picture on the lid of the box would be of a crowd of happy library users, looking, listening and playing with digital objects which their parents and grandparents created, but rendered in their own computer environment. When this picture becomes a reality, it will demonstrate that the library community preserved the heritage in the right way and that it guaranteed the accessibility and usability over the years.

In the past few years, much effort has been devoted to raising awareness of the issue of digital preservation, especially amongst cultural heritage institutions. All those articles, presentations and discussions are gradually beginning to pay off. Digital preservation is no longer a topic that needs to be explained. On the other hand, however, the ultimate goal, the picture on the lid of the box where all the different pieces will become a coherent entity, is still not a reality. Although we are making progress, work is too fragmented and has not led to an out-of-the-box solution. Lots of people are working in the area of digital preservation, but still much effort is needed to integrate the work done on separate pieces of the puzzle.

There are lots of methodologies to complete a puzzle. Some people start by looking for the corner pieces, other people will complete the outer edges first and yet another category will first collect the blue and white pieces to finish the clouds. In digital preservation similar processes are taking place. With so many organisations involved, the list of topics related to digital preservation research gets longer every day. In this article I will make a selection and show you the current state of affairs in three areas:

- the place of digital preservation within an organisation;
- developments with regard to the digital objects; and
- the effects of international collaboration.

The Place of Digital Preservation within an Organisation

Digital preservation is an intrinsic process and not a separate activity. Whether it concerns a library or an archive, digital preservation affects the organisation as a whole and should not be an isolated activity. Work flows need to be designed for collection policies and management, selection and appraisal, metadata, access procedures etc. — in the same vein as for printed collections.

Several initiatives have been devised to support organisations in implementing digital preservation, both for newcomers such as organisations starting to think about setting up a digital repository and more experienced organisations wishing to evaluate their policies and the effects of their preservation activities.

(Self) Auditing

The status of being a trusted, or more correctly termed a *trustworthy* repository is the ultimate goal of an organisation with a digital collection that needs to be accessible and usable over time. A first initiative designed to raise the issue of certification and to provide guidelines for a trusted repository was the joint publication in 2002 by the US Research Libraries Group (RLG) and OCLC of *Trusted Digital Repositories: Attributes and Responsibilities*. Many organisations are presently using this document as a checklist. The success of this document and the need for a real auditing instrument led in 2007 to a new initiative designed to update these guidelines with the latest insights and experiences, and to turn it into a clear, understandable ISO standard which can be used as a certification and auditing instrument in the digital preservation community. To involve as many parties as possible, everyone interested can participate in this initiative; the discussions and outcomes are publicly available.[1] It can often take years to create an ISO standard, but a first draft will be available by the end of this year.

Audit and certification can also be looked at from a different angle, as is done by the DRAMBORA initiative. This 'Digital Repository Audit Method based on Risk Assessment' looks upon digital preservation as the task of managing risks. It offers training and tools to perform a risk analysis of the organisation in order to identify areas that can be improved. A third initiative is the *Catalogue of Criteria for Trusted Digital Repositories* (2007) by the German nestor group.

The three initiatives mentioned cooperate closely, and in 2007 they jointly formulated the ten core principles of trust[2] as leading principles for trustworthy repositories. These ten core principles were used as input for the PLATTER tool (Planning Tool for Trusted Electronic Repositories),[3] specifically developed to help organisations starting with digital preservation programmes to implement these principles and be able to meet the audit and certification requirements. To achieve this, trained and skilled staff is needed who constantly update their knowledge. Several European projects on digital preservation, such as DPE, PLANETS (Preservation and Long-term Access through

Networked Services) and CASPAR (I will discuss these below), have specifically mentioned dissemination of knowledge as one of their deliverables and they offer training by experts who update staff on the latest insights in various aspects of digital preservation, often in joint workshops.

The cost of digital preservation still is an interesting and very important topic. As digital objects cannot be ignored, even for a while, at any stage during their life cycle, insight in costing required for the long term, is vital. One of the major initiatives in this area is the LIFE project, LIFEcycle Information for E-literature, a collaboration between University College London (UCL) Library Services and the British Library, which is funded by the Joint Information Systems Committee (JISC). The first stage of this project resulted in the development of a costing model for the different processes taking place in the life of a digital object. Starting with creation, and on to preservation, to access and usability, every activity in these processes involves costs for the preserving organisation, such as acquisition activities, metadata creation, storage, but also preservation watch and preservation action. In 2007–2008 a second iteration of this LIFE project was funded by JISC. This phase will lead to an economic evaluation of the model and an update based on the results of several cases studies involving different kinds of digital material.

Related to the question of costs is that of the 'value' of collections. How do we value a digital collection? What material does a library need to preserve? For example, if a library digitises part of its collection, does it need to preserve the digital master files for the long term, or is it more economical to preserve the paper collection and maybe digitise it again sometime in the future? And how about a full domain crawl of the national websites? As websites are growing every day, it is a huge task for a national library to organise a representative domain crawl. The technical means to implement selections in a full domain harvest are limited. On the other hand storage costs might be a reason to make choices and to select. Such a selection is one of the topics the European LiWA (Living Web Archives) project will focus on, but the topic of appraisal and selection is also frequently mentioned in conferences and articles.[4]

One of the aspects of digital preservation that is not solved yet, is rights management. When preserving digital material, it might be necessary to perform actions on the digital objects in order to keep the object accessible and usable. These actions might conflict with copyright laws. Preserving organisations are not always sure if they are allowed to perform the necessary tasks. Is it allowed to make multiple copies of a work for preservation purposes? Or to migrate works to a new technological format, thus creating a new manifestation of the original object? National laws are often not updated for the digital age, and if they are, this aspect is regularly left unresolved. Recently a study[5] drew attention to this problem; in conclusion it presented a set of joint recommendations to provide guidelines for national copyright and related laws.

Digital Objects

We looked into the organisational aspects and the trends in that area, but what about the digital objects themselves? Do they change in a technical sense? For a long time the majority of digital objects were rather straightforward, often consisting of one file in a well-known format like PDF or TIFF. Many digitisation programmes resulted in large quantities of objects in TIFF format. But the digital world is getting more complicated, the users are changing and becoming more demanding, and this is reflected in the digital objects themselves. Websites are a well known example, as the sites become more complex and offer more features. Long-term archiving of the results of domain harvests is a topic even the International Internet Preservation Consortium (IIPC) is slowly taking up, focusing more on harvesting itself than on the long-term archiving aspect.

There is also a tendency to link publications with data bases, websites, blogs etc. to offer the end user a single point of entry to all related publications. This is especially true in the world of institutional repositories, but academic publishers also increasingly allow authors to include other types of digital material within their article. As a memory institution you might want to preserve this set of materials and offer your future users access to it. But the various components of this package might not be located in the same repository. The European DRIVER project will investigate the consequences of these so-called enhanced publications for long-term preservation. One of the essential requirements to preserve this material will be the use of persistent unique identifiers to accompany the publications during their entire lifetime. Another requirement will be interoperability between objects in different repositories, using standards for interoperability. These developments will not only be interesting for institutional repositories, but, as the boundaries between a publication and the linked digital attachments become more blurred, it will become important for national (deposit) libraries as well.

As accessibility and usability of digital objects are the principal goals of digital preservation, it is important to gather all of the essential information needed to render the object correctly in the future. Apart from file format and version information, you want to have information on other aspects like behaviour and appearance of the object at the time it was created — in other words, the 'significant properties' of an object. For reasons of economics and efficiency, this information should be collected automatically. Several reference services to collect these kinds of information are already available in a basic form, but they are presently being updated: the PRONOM registry of file format information of the National Archives in London will be expanded in the PLANETS project; the JHOVE project, comprising tools to validate and characterize file formats, received new financial support to start JHOVE2, the UK InSPECT project published some interesting studies, and international initiatives have been taken to set up a Global Digital Format Registry (GDFR). Supporting tools were also built elsewhere, such as the Metadata Extraction Tool of the National Library of New Zealand and the XENA tool of the National Archives of Australia which normalises various file formats.

Although all of these initiatives are warmly welcomed by the preservation community, they do have one major drawback: lack of sustainability. Nearly all information about digital preservation that has been generated by all these projects is freely available from the internet and tools can be downloaded at no cost at all. But there is a risk that these supportive tools for digital preservation will not be managed properly after the projects are finished. Therefore, although there is enthusiasm about the initiaitives, organisations are hesitant to rely on these tools and build their own, sometimes unnecessarily.

The topic of sustainability is especially important in relation to the results expected from the major European projects PLANETS and CASPAR. Who will maintain the tools developed? Who will update and monitor the information of file format registries that so many preserving organisations will rely on? If this topic of sustainability is not solved, a lot of effort will be wasted. The solution might be found in my last topic, that of international collaboration.

International Collaboration

The 6th and 7th Framework programmes of the EC generated a number of projects with a focus on different aspects of digital preservation. Three important projects: PLANETS, CASPAR and DPE are now half way and are beginning to present their (intermediate) results at conferences and on their websites. In the PLANETS project, with the main focus on preservation planning, the PLATO tool is taking shape. This decision support tool will help an organisation to plan its preservation activities. In two years time this supportive tool will be integrated with a test bed (where you can perform tests with samples of your collection) and registries with information on preservation tools which you can use for preservation actions. As I said, the further development of the PRONOM registry with file format information will be part of this project. Digital Preservation Europe (DPE) was funded to bring digital preservation expertise together and to develop a roadmap for future research in the area of digital preservation.The project also published practical solutions like the aforementioned audit tool DRAMBORA and the PLATTER tool.

A new project is KEEP (Keeping Emulation Environments Portable), in which emulation as a preservation action will be developed further and will be integrated into a framework for use in the preservation community. KEEP will follow on from the emulation development work done by the Koninklijke Bibliotheek, National Library of the Netherlands in collaboration with the National Archives of the Netherlands, which resulted in 2007 in the launch of the DIOSCURI emulator. DIOSCURI will be further developed within PLANETS. The KEEP project will help to put DIOSCURI in a broader context with other emulator tools.

The European projects also focus on other areas, such as innovative storage methods in the CASPAR project. The SHAMAN project (Sustaining Access through Multivalent Heritage Archiving) focuses on different aspects of digital archiving systems and, as mentioned before, the LiWA (Living Web Archives) project deals with websites.

Several European national libraries are participating in these projects and contribute with both practical as well as professional knowledge. Participating research institutes and commercial partners have their own skills and are often more experienced in IT-related areas, which make them important partners in furthering digital preservation. This mix of participants is crucial for the success and acceptance of the project results.

In Conclusion

I have given you an overview of the latest developments in digital preservation, with a focus on organisational aspects, the digital object itself and progress in EC co-funded projects. Collaboration in digital preservation is crucial, as is often mentioned. Initiatives on a larger scale, like the European Alliance for Permanent Access, should help to unite the scattered pieces and to complete the jigsaw puzzle of digital preservation.

Websites Referred to in the Text

Alliance for Permanent Access, http://www.alliancepermanentaccess.eu

CASPAR, http://www.casparpreserves.eu/

DIOSCURI, http://dioscuri.sourceforge.net/

DPE, Digital Preservation Europe, http://www.digitalpreservationeurope.eu/

DRAMBORA, http://www.repositoryaudit.eu/

DRIVER, Digital Repository Infrastructure Vision for European Research, http://www.driver-community.eu/

GDFR, Global Digital Format Registry, http://www.gdfr.info/

IIPC, International Internet Preservation Consortium, http://www.netpreserve.org/about/index.php

InSPECT, http://www.significantproperties.org.uk/

JHOVE2, http://confluence.ucop.edu/display/JHOVE2Info/Home;jsessionid=6A3E8B4924066A596523FF0F3127C5EF

LIFE, Lifecycle information for e-literature, http://www.life.ac.uk/

LiWA, Living Web Archives, http://www.liwa-project.eu/

nestor, http://www.langzeitarchivierung.de

PLANETS, Preservation and Long-term Access through Networked Services, http://www.planets-project.eu/

PRONOM, http://www.nationalarchives.gov.uk/pronom/

SHAMAN, Sustaining Heritage Access through Multivalent Archiving, http://shaman-ip.eu/

Notes

[1] All links were checked on 5 September 2008. See http://wiki.digitalrepositoryauditandcertification.org/bin/view/Main/WebHome

[2] These principles are described in the "DPE Repository Planning Checklist and Guidance DPED3.2", p. 9, http://www.digitalpreservationeurope.eu/publications/reports/Repository_Planning_Checklist_and_Guidance.pdf

[3] Published in 2008, see note 2.

[4] See: S. Ross (2007), *Digital Preservation, Archival Science and Methodological Foundations for Digital Libraries*, Keynote Address at the 11th European Conference on Digital Libraries (ECDL), Budapest (17 September 2007).

[5] International Study on the Impact of Copyright Law on Digital Preservation. A Joint Report of The Library of Congress National Digital Information Infrastructure and Preservation Program, the Joint Information Systems Committee, the Open Access to Knowledge (OAK) Law Project and the SURFfoundation, 2008, http://www.digitalpreservation.gov/partners/resources/pubs/digital_preservation_final_report2008.pdf