IFLA WLIC 2017 WROCŁAW

Submitted on: 26.06.2017

Preserving digital legal deposit - new challenges and opportunities

Raivo Ruusalepp
Director of Development, National Library of Estonia, Tallinn, Estonia.
E-mail address: Raivo.Ruusalepp@nlib.ee

Abstract:

Born digital content has always been considered to be a bigger challenge for preservation than digitised content. Higher volume and technical complexity, dynamism as well as a complex surrounding rights space are frequently cited as aspects that make born digital content 'special' to memory institutions. This paper builds on the Estonian case of introducing digital legal deposit which has led to an exercise of reconceptualising the digital preservation function of the national library. The rapid increase in volume, file size and new file formats have led to making the library's preservation service levels explicit, an update to the preservation policy and automation of archiving workflows. The new demands on preservation are pushing the current digital repository system of the national library to its limits and the library needs to embark on migrating to a new preservation solution. This response to a sudden change in digital preservation workload is typical in the heritage sector – upgrading the ingest component is the first instinctive reaction of most memory institutions. This paper proposes that increasing the throughput of ingest component needs to be combined with a modular concept of a preservation system that sets interoperability as its core principle. When digital preservation is conceptualised as an exercise of resilience rather than sustainability, the interoperability requirement for systems architecture and service design follows logically.

Keywords: digital legal deposit, preservation of born digital content, resilience.

Introduction

In 2010 an OCLC survey report concluded that "management of born-digital archival materials is still in its infancy" (Dooley, Luce, 2010). But already five years later, the American Libraries magazine called for help: "We are in trouble! The scope, depth, and cost of the threat mean that individual libraries cannot advance born-digital content preservation on their own." (Neil, 2015). The sheer volume of data that libraries should be archiving has grown rapidly and can no longer be handled on a one-by-one basis, as can analogue materials or as digitised content has often been processed for preservation.

Introducing digital legal deposit

National Library of Estonia (NLE) first started collecting digital print files from public agencies and newspaper publishers in 2006. The files were stored in the Fedora Commons repository system, freshly installed and configured as part of the European reUSE project (https://www.uibk.ac.at/reuse/). As the number of publishers agreeing to deposit their print files with the library grew, a standard contract was developed that fixed the responsibilities of both parties similar to the legal deposit procedure. An archiving modules were developed for the repository solution for both the depositors and librarians managing the collection to support the what was effectively voluntary digital legal deposit. By 2014, all newspapers published in the country were deposited also digitally to the National Library. This became a tipping point for starting to revise the existing Legal Deposit Act. The ensuing debate with publishers over changes to the current working model of legal deposit was occasionally emotional but constructive and not overly prolonged. Ten years after first born digital print files were archived with the National Library, the Estonian legislature passed a new Legal Deposit Copy Act (Legal Deposit, 2016).

So from January 2017, Estonia joined the handful of countries where legal deposit encompasses also digital versions of publications. The new legal act put an explicit focus on preservation (in Estonian the act is literally called 'Preservation Copy Act') and stipulates how the production formats of printed publications must be deposited with the NLE, and production formats of films that have received state subsidies must be deposited with the National Archives' film archive. The Act also covers web archiving of the national domain as the mandate of the NLE.

The many concerns of publishers that departing from the print files used to produce their publications will damage their commercial interests were not always easy to address and to establish continuing trustworthiness of the National Library it was necessary to be very transparent (sometimes down to personal visits – "seeing is believing") about the archiving, storage and preservation processes at the library. The commercial risks and concerns had to be balanced against authors' expectations towards publishers about a permanent archive of their work. The National Library took on the role of a permanent national digital archive of publications where the publishers can always retrieve their print files in their original format for free. Thus, the NLE became part of the value chain in a longer publishing cycle as a service provider. In return, NLE's digital preservation remit became somewhat more complex.

The first five months of digital legal deposit

The implementation of the new Legal Deposit Copy Act has started smoothly. A new e-services portal was developed for publishers that is also used for depositing print files and publishers have adopted the new tool with gusto.

In terms of digital preservation, the most visible result thus far is that the volume of deposited files has grown by around 70% compared to the previous years. The average object size has increased by some 30% and this, combined with the growth in overall volume, requires rapid response in enlarging the storage space available at the library. We also detect a greater variety of file formats and a larger proportion of files with long term preservation issues that our initial quality control detects at ingest, for example fonts not embedded in PDF, epub conformity to standard, etc. Based on user satisfaction surveys that we run regularly among

publishers and users of the library's digital collection we also see that the user expectations for how fast the library can process the deposited files have become more demanding – minutes if not seconds are the norm now, and the underlying assumption is that the file processing workflow from publisher to the user interface is fully automated.

The library's response to the new preservation challenges

The changes in volume and complexity of born digital content was expected and so the National Library put measures in place to anticipate the new demands. These included updating significantly the digital preservation policy; developing and publishing a new recommended file format list for deposit and preservation; explicit definition of service levels based on the NDSA matrix of digital preservation services (NDSA, 2013).

Based on the few months of practice, the stages in the workflow with most errors that require human intervention have been identified and a next stage of software development is commencing to re-design and automate further workflows around the ingest and archiving. One of the new aims is check the quality of submitted files as early as possible in the process and make the results available also for publishers, together with information on implications that imperfect files will have on subsequent preservation levels at the library.

Despite all this, we find ourselves again in a situation that was aptly described by publishers already almost 20 years ago:

> Publishers do not undertake the preparation of their resources with a view to long-term preservation, but rather with a view to making the product as attractive as possible to its potential market, aiming to maximise both quality and sales. This may involve the use of formats that will not prove easy to preserve in the long term, and the deposit libraries cannot recommend that the resources are prepared in other, "preservation-friendly" formats instead (or rather, they can make recommendations, but they cannot expect that publishers will necessarily follow them). (Bide, Potter & Watkinson, 1999)

Is this response typical of memory institutions?

Has the natural response of the NLE been typical of how memory institutions react to the rise in volume and complexity?

When faced with challenges of scale, preservation institutions tend to focus on automation. The digital preservation community has been striving towards speeding up and automating the workflow for more than a decade now but has not really moved beyond semi-automation. By and large, most institutions are still solving ingest problems, NLE among them. As part of this, memory institutions are attempting to push the responsibility 'upstream', towards the creators of objects and thus share some of the responsibility and workload of preservation. Examples like research data management plans that are made a condition of funding by research funding agencies; the BagIt container for depositing digital materials to a repository. Giving an early warning that something is not fit for preservation or that it needs action now is part of this process and the recent Preforma project has produced excellent tools for facilitating this (Preforma, 2017).

Complexity poses questions about what the object of preservation really is and manifests itself mostly as a rendering problem – how can we provide access to a complete object when we ourselves only hold part of it? For example, smartphone apps for on-line news, newspapers and magazines that legal deposit libraries are finding difficult to ingest and preserve because of technical complexity, proprietary technology and high cost. Limiting the scope of service that a memory institution can provide is a typical response, by connecting preservation service levels with file formats that are known to the preserving institution.

Can these responses offer a longer term solution or are they mere quick fixes that help remedy acute pain in some parts of the system? Is ingest the only bottleneck in the process? Or is a more substantial reconceptualization of the entire digital preservation exercise required?

Redefining digital preservation at the NLE

The National Library started to develop a new vision for its digital preservation service already before the digital legal deposit was enacted. Anticipating a step-up in volume of objects to be ingested, the workflow modules of the digital archive solution were critically assessed and their performance deemed poor. Larger volumes of objects require new tools to present the material in more manageable forms that go beyond the usual click-and-view or click-and-download interfaces. In Estonia, memory institutions fall under the legislation that requires open licence data to be made available as open, preferably linked data, and provide text mining services for academic research. The current preservation infrastructure at NLE is not up to providing the users with options to mine or manipulate the data. Technical capabilities of the digital repository system and its ability to provide APIs (application programming interface) in support of new types of services form only one aspect of the problem. A much more challenging issue is the conceptual structure of the objects that the library preserves – the archival information package (AIP; see: ISO 14721) has to be adaptable for new types of content and new services based on its content. Resilience to change is a requirement for AIPs and this includes migration to a new repository system.

From the very beginning of discussions around digital preservation it has been clear that the core issue is that information needs to live longer than the system(s) that produced it. The same applies for digital preservation systems and digital repositories. William Kilbride summarised this in a recent Digital Preservation Coalition blog (Kilbride, 2017): “We need to embrace our own mortality: digital preservation tools are products of their own times so are a contingent solution to an enduring problem.” Digital preservation systems are subject to the same kind of obsolescence that they are designed to prevent.

In its design and planning for a new digital preservation solution NLE is facing the challenge of migrating the contents of its existing digital repository to a new solution, while keeping the existing services built on the content running and being able to develop new services, like text mining and machine-to-machine queries via linked data protocols. This is a challenge that departs from the level of file formats that sometimes seems to have become the focal issue of digital preservation. When digital preservation is looked at systems level, other aspects come to focus, especially interoperability.

Resilience as a concept for preservation systems

Digital preservation discourse has for a long time been focussed on longevity and sustainability issues of file formats, workflows, tools, etc. This has resulted in a plethora of tools and services that are tailored for a specific object type or support a particular task or require a specialised skill-set to implement. Not only has it left preservationists, especially those new to the discipline, perplexed about what works best for what preservation situation, it has also diverted the attention away from the systems level.

When defined as a systems-level issue, digital preservation becomes largely an interoperability challenge. For example, content objects and services based on them in one repository system can be migrated to a new repository system by defining interoperability requirements and standards that support them. An example of this approach is offered by the recent eARK project (http://www.eark-project.com/) that took interoperability as the binding concept for the preservation toolset it developed and has established an Archival Standards Board (http://www.dasboard.eu/) to ensure longevity of the standards that support the interoperability.

Standards-based interoperability as a means of overcoming obsolescence of systems brings resilience into spotlight of preservation. Rather than longevity or sustainability of digital information, the next stage of maturity of digital preservation domain should focus on resilience of systems and information created by them. Resilience in this context can be defined as the capacity to prepare for disruptions, recover from shocks and stresses and adapt and grow from a disruptive experience. The first generation of digital repositories and digital preservation systems are reaching an age where they can be categorised as legacy software. The process of replacing them will be undertaken by many memory and academic institutions in the few coming years. Defining resilience conditions for preservation systems as interoperability requirements when migrating between systems, would help to conceptualise digital preservation in new ways and make this domain more future proof.

Conclusion

A quarter of century has passed since the widespread use of personal computers and creation of the world wide web, and very large wave of born-digital content is awaiting to reach preservation institutions for archiving in near future. Memory institutions around the world are preparing for this arrival by learning to cope with the bigger scale and increasing complexity. However, the volume of content is such that the solution cannot be only in better ingest tools or organisational measures like adding manpower for processing files. A rethink of what digital preservation is for and how it can be achieved is also required in order to learn to live with constant change. Focussing on resilience of systems through standards-based interoperability is a useful way forward. Starting to learn from migrating our own legacy preservation systems will allow us to teach our information producers what preservation-aware systems look like in the future.

References

Bide, M., Potter, L., Watkinson, A. (1999). Digital preservation - An introduction to the standards issues surrounding the deposit of non-print publications. British Library and Information Commission Research Report 23. Retrieved from www.bic.org.uk/files/pdfs/digpres.pdf

Dooley, J.M., Luce, K. (2010). Taking our pulse: The OCLC Research survey of special collections and archives. Dublin, Ohio: OCLC Research. Retrieved from http://www.oclc.org/research/publications/library/2010/2010-11.pdf

Estonian Legal Deposit Copy Act. (2016). Retrieved from https://www.riigiteataja.ee/en/eli/514092016001/consolide

ISO 14721:2012 Space data and information transfer systems – Open archival information system (OAIS) – Reference model

Kilbride, W. (2017). Obsolescence 2.0 Digital Preservation by people, for people. Retrieved from http://dpconline.org/blog/obsolescence-2-0-digital-preservation-by-people-for-people

National Digital Stewardship Alliance. (2013). Levels of Digital Preservation. Retrieved from http://ndsa.org/activities/levels-of-digital-preservation/

Neal, J.G. (2015). Preserving the Born-Digital Record. Many more questions than answers. Retrieved from https://americanlibrariesmagazine.org/2015/05/28/preserving-the-born-digital-record/

Preforma. (2017). Open source portal: veraPDF, DPF Manager, Mediaconch. Retrieved from http://preforma-project.eu/open-source-portal.html