Towards Preserving Cultural Heritage of Finland

Heikki Helin, Kimmo Koivunen, Juha Lehtonen, Kuisma Lehtonen
CSC – IT Center for Science
P.O.Box 405, FI-02101 Espoo, Finland
firstname.lastname@csc.fi

We present the status and challenges in digital preservation of the National Digital Library (NDL) of Finland. The NDL aims to create a nationally unified structure for contents and services ensuring the effective and high-quality management, dissemination, and digital preservation of cultural digital information resources. NDL's basis is formed by libraries, archives, museums, and other organizations in Finland storing cultural heritage material and the actors responsible for their IT solutions affecting hundreds of organizations. We believe that nationally shared digital preservation infrastructure and services will draw the practices of memory organizations closer, reduce the costs and fragmented nature of the systems, and intensify cooperation.

Given the diversity of partner organizations, the digital content to be preserved makes up a very heterogeneous whole. A major share of content owned and administered by partner organizations consists of digitized documents and photographs, but the volume of born-digital content is expanding quickly. In the future, the largest content categories of born-digital content in the NDL will be recorded TV and radio programs, films, and digital documents and publications. Based on extensive surveys conducted among partner organizations, we estimate that digital information stored to our digital preservation system by 2020 will consists of more than 2 400 million objects requiring more than 12 petabytes without replication.

The digital preservation system will be built to accommodate the increased volume and diversification of content and organizations, as well as the possible development into a storage system for the preservation of research data. A quick launch of the bit preservation will ensure that the digital information in the possession of the partner organizations can be reliably preserved until the system becomes fully operational later, including trustworthy digital preservation.

Keywords: cross-sector collaboration, cultural heritage material, digital preservation, scalability

Introduction

Digital preservation (DP) refers to the reliable preservation of digital information for several decades or even centuries. Hardware, software and file formats will be outdated, while the information must be preserved. In bit preservation, we ensure that the actual bits remain intact and accessible at all times, is certainly the basic requirement, but also the authenticity and interpretation need to be preserved, and the content needs to be kept understandable.

Additionally, in some cases, the original user experience evoked from the data needs to be conveyed for the future users. Therefore, digital preservation is not just a technical challenge, but it also requires skills with operational, cognitive, financial, and legal capabilities to be managed successfully. As a result, a plan is needed for preservation actions required to keep the content intact, authentic and to ensure the accessibility and reliability of the data.

The National Digital Library (NDL) of Finland is an entity within the remit of the Ministry of Education and Culture. NDL's basis is formed by libraries, archives, museums, other organizations that store cultural heritage material and actors responsible for their IT solutions (partner organisations). In addition to the DP system, the aim of the NDL is to ensure the effective and high-quality management, dissemination, and centralized digital preservation of cultural and scientific digital information resources. The objectives of the NDL are ensuring the preservation of digital cultural content, ensuring access to and compatibility of content, designing a cost-effective DP system, promoting cooperation between the partner organizations, and building better services with open cooperation and expansion to include a large range of content. Centralized and shared infrastructure and services will draw the practices of memory organizations closer, reduce the costs and fragmented nature of the systems, and intensify cooperation.

In the NDL, the key, prioritised data objects will be digitised and made available through the public interface (see Figure 1). The partner organisation will manage its own content information through its back end systems, from which information will be harvested to the public interface. User searches will be directed to the indexed aggregated database of the public interface and, where necessary, the digital object obtained from the back end system for access by the user.

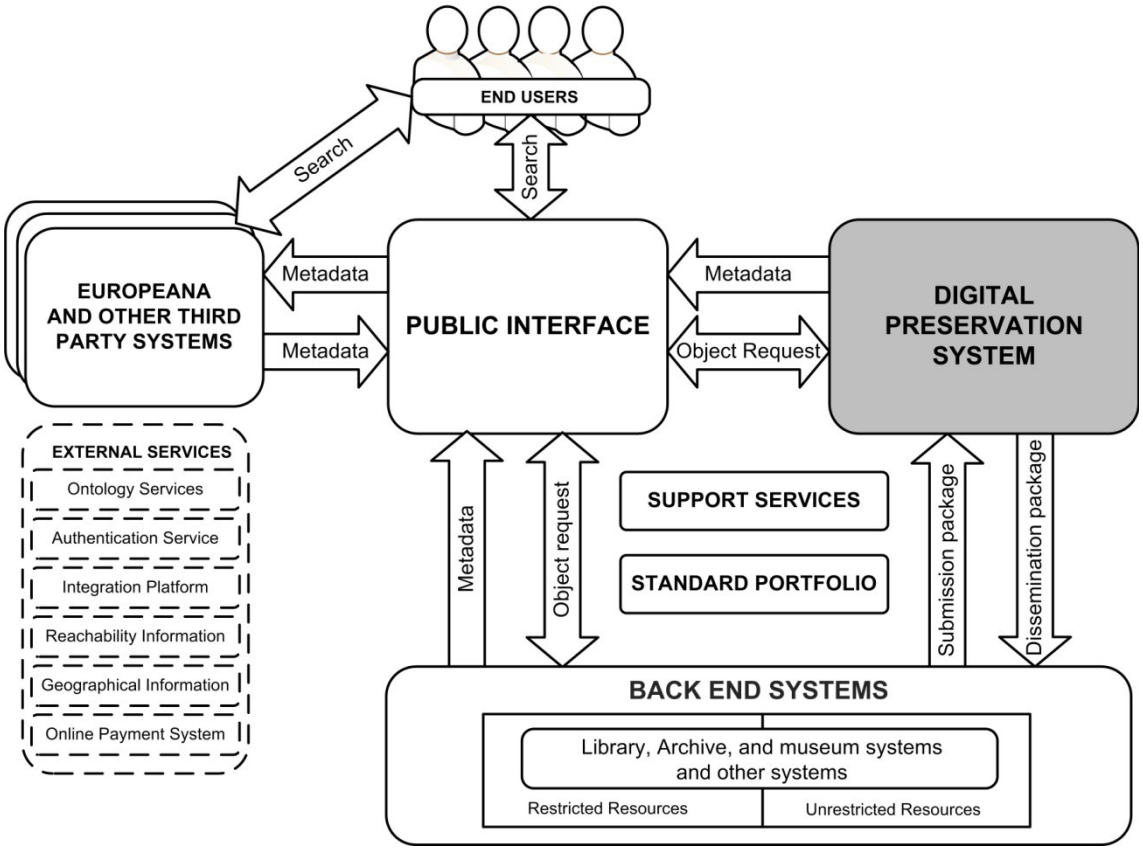

Figure 1: General overview of the National Digital Library

CSC – IT Center for Science is currently responsible for the design and implementation of the NDL's DP system. In this paper, we present an overview of this system including the current status and challenges. Our DP system will be based on the OAIS reference model [1]. The implementation is divided into two main phases: the preparation and implementation. The preparation phase will ensure that the original bits of the data can be maintained intact and run on modern hardware. A quick launch in 2013 of the bit preservation will ensure that the digital materials in the possession of the partner organisations can be reliably preserved until the DP system becomes fully operational in 2016. In the second phase, the DP system implementation will ensure that the material remains understandable, its information content can be interpreted, and the material can also be used with the software of coming decades. The system will be built to accommodate the increased volume and diversification of content and organizations, as well as the possible development into a storage system for the preservation of research data.

Data Volume

Almost all memory organizations under the Ministry of Education and Culture of Finland are under an obligation to preserve a lot of their digital material. Most of the cultural heritage material consists of digitized documents, maps, photographs, newspapers and sound recordings. In the future, this material will be mostly born-digital, which increases the volume of the data.

To support NDL's DP planning, we carried out a survey in Spring 2011 with the partner organizations to clarify the scope and various other aspects of the digital information resources managed by the partner organisations that will be stored in the NDL's DP system. The survey helped to specify the extent of the digital material of partner organisations and, in particular, it examined what kind of preservation methods they require. The survey results are summarized in Table 1. It can be seen that the data volume increases rapidly and that the space requirement with different types of objects is very heterogeneous.

Table 1: Data volume estimates in the 2011 survey.

	2010		2011		2015		2020	
	Number of objects (millions)	Size (TB)	Number of objects (millions	Size (TB)	Number of objects (millions	Size (TB)	Number of objects (millions	Size (TB)
Files and documents	11.6	328	15.4	394	25.6	646	48.7	1,301
Photos	1.7	18	2.1	30	3.9	68	6.1	120
Films	0.1	495	0.2	1,143	0.8	3,055	1.2	8,020
Sound recordings	1.2	606	1.5	771	2.4	1418	3.7	2,176
References	19.5	1.2	21	1.5	27	2.4	34	3.4
Online archive	496	20	646	27	1,396	59	2,300	97

Radio and TV archive	0.8	95	1.2	142	2.9	327	5.0	558
Total	**530**	**1,563**	**687**	**2,509**	**1,458**	**5,575**	**2,400**	**12,275**

The data volume in NDL is large and includes lots of large objects. Amount of data to be preserved will cause a lot of scalability issues that have to be solved before the system can be released to production. Different processes, such as ingestions, integrity monitoring, refreshments, replications to different media (e.g., tape data storage) and migrations require resources. The DP system needs to be built with an idea of local distributed storage systems. In our system, the storage nodes will be low-cost resource servers independent of each other, each taking care of a local RAID6 disk array. The servers will be bonded with a redundant file catalog database. This kind of distributed system is very scalable: The system can give free resources for different processes automatically, and the total costs caused by adding or upgrading the hardware piece by piece remain at a reasonable level. Of course, other media types are also needed, such as tape library for the backup purposes. Additionally, when building locally distributed system to separate geographical sites, the network capabilities and reliability must be considered very carefully.

Digital Preservation Services

The selection of digital information and its preparation for digital preservation will primarily be the responsibility of the partner organisations. However, number of services to facilitate the work will be offered by the DP system. These services include pre-ingest, ingest, archival storage, usage, management and support services.

The specifications prepared in the NDL will guide both the preparation of the digital information in partner organisations and the ingestion activities of the DP system (such as validating the digital information). Adequate and consistent metadata will be essential for the success of digital preservation. A generic METS [2] profile has been specified within NDL describing common features and their application instructions for all submission information packages (SIP), including the mandatory and recommended metadata elements. The specification will take into account both semantic unity and technical implementation, and it will be updated along with progress in the DP planning process. At a later date, sub-profiles will be defined for different content types that further define the characteristics associated with the information in question. This will ensure that the SIP contains all the necessary information to enable the digital preservation. SIPs will be automatically validated within the DP system before they are formed into archival information packages (AIP). The inspection process and conversion may require a lot of resources, and therefore it is important to distribute the process concurrently to different low-cost servers.

Each of the objects transferred to storage must have a pre-defined preservation plan which defines the objectives of DP and the methods employed to achieve these objectives. The understandable representation capability of the digital information can be ensured with following the plan. Preservation plan templates will be drafted for common types of digital information that can be applied when the preservation objectives and methods are similar. Preservation planning also includes the monitoring and updating of plans throughout digital information life cycle.

Following the report by Library and Archives Canada [3], two types of file formats that the NDL DP will be supporting are identified as recommended or acceptable for transfer.

Recommended file formats are ones for which the NDL project foresees a long useful life (e.g., PDF/A), and the data submitted in these formats will be accepted into storage as it is. The acceptable for transfer file formats are those in which a lot of digital information for digital preservation is stored in the NDL (e.g., various MS office formats), but which may be inappropriate as a long-term solution. The objects submitted in acceptable for transfer file formats will be migrated into a recommended format before storing those to the DP system.

The NDL will also provide a packaging service application, which helps the partner organisations to form a SIP compliant with the NDL METS profile. The packaging service will verify that all mandatory metadata of each object has been added and will then independently form the SIP. This packaging service application is designed especially for small partner organizations, which may have only a relatively small amount of information to be preserved and thus do not have resources to implement interface to their back-end system supporting the DP system, or, as in many cases, they lack the back-end system completely.

The functionality and services relating to the preservation of digital information are the key components of the DP system. This will guarantee the reliability, understandability, and immutability of the material. The preservation plan [4] describes, for example, how the AIPs are copied to various types of media and how the copies are managed. As recording equipment becomes outdated, copies will be transferred onto new, modern hardware and storage media. Data reliability will be ensured through appropriate security arrangements.

The DP system will ensure that the original bits of the digital object remain unchanged and can be run on modern hardware. Multiple copies of the digital object and its metadata will be created to different types of media and stored in (at least) two different geographical locations. In our current plan, these sites are located in Espoo and in Kajaani, Finland, with about 550 km distance between. The connection between the sites will be done with a fiber backbone. The integrity of the copies will be monitored at regular intervals with checksum calculation processes. Since the data volume is high and includes large objects, the integrity monitoring process needs to be designed as a parallel operation for several servers. AIPs stored in the DP system will be refreshed at regular intervals in line with the OAIS reference model [1]. Preserving all essential metadata together with the digital objects in the DP system will keep the material understandable. In addition, the digital object will be migrated into the format used at any particular time in the future. Migrations will be designed and tested in collaboration with partner organizations in line with the OAIS model.

The DP system will include search functions necessary to find content. The digital information in storage can be searched using tags or keywords. By default, only the partner organisation that has ingested the data can collect the required version of the digital information transferred into storage from the DP system in situations where the material has been damaged within the organisation’s own system or when its authenticity is uncertain. Licenses and access rights determine which other parties can access and use the content.

In line with the OAIS reference model, the DP system will deliver the digital information as dissemination information packages (DIP) that are formed in compliance with common specifications. DIP formed by the DP system will also be formed in accordance with the NDL METS profile. The DIP does not necessarily have to retain the actual preserved digital objects, but it may only contain metadata.

Data management services include functions that, for example, allow the partner organisation to upgrade the metadata of the preserved digital information and add new versions of the

content. Digital information may also be removed from storage if necessary. When digital information is removed, this will be recorded in the system on what was removed and why. In addition, a variety of reports and statistics on the digital information, its use, and completed preservation actions can be retrieved. Data management services will also include the updating of preservation plans, DP system maintenance, cost control, and risk management.

The DP system will comprise a series of complementary advisory and support services, suitable for different situations and organisations of different sizes. These will include for example support for the use of the DP system, training sessions or administrative support for the partner organisations.

Discussion

Shared digital preservation infrastructure and services will draw the practices of partner organizations closer, reduce the costs and fragmented nature of the systems, and intensify cooperation. However, the heterogeneity of the data with several partner organizations gives certain challenges to specifications. Although a common specification (profile) already exists, it most likely will be updated several times in the future. Since the needs of all partner organizations must be considered, the specification update process requires a lot of discussion and collaboration.

Acknowledgements

The authors would like to thank all members of the NDL digital preservation support group and technical division for their valuable comments and input during the preparation of the NDL digital preservation system. Further, we thank the Ministry of Education and Culture for funding this project.

References

ISO: Open Archival Information System—Reference Model (ISO 14721:2003). International Standards Organization, 2003

METS Metadata Encoding and Transmission Standard: http://www.loc.gov/standards/mets/

Library and Archives Canada. Local Digital Format Registry (LDFR): File Format Guidelines for Preservation and Long-term Access, 2010

Christoph Becker, Hannes Kulovits, Mark Guttenbrunner, Stephan Strodl, Andreas Rauber, and Hans Hofman. Systematic planning for digital preservation: evaluating potential strategies and building preservation plans. International Journal on Digital Libraries, Volume 10, Issue 4 , pp 133–157, 2009