On Preparedness of Memory Organizations for Ingesting Data

Juha Lehtonen, Heikki Helin, Kimmo Koivunen, Kuisma Lehtonen
CSC – IT Center for Science
P.O. Box 405 (Keilaranta 14)
FI-02101 Espoo, Finland
+358 9 457 2001
{juha.lehtonen, heikki.helin, kimmo.koivunen, kuisma.lehtonen}@csc.fi

ABSTRACT
The National Digital Library of Finland is an entity, which aims to create a nationally unified structure for contents and services ensuring the effective and high-quality management, dissemination, and especially digital preservation of cultural digital information resources. The National Digital Library's basis is formed by libraries, archives and museums (partner organizations). Because of the diversity of the partner organizations, the digital content to be preserved makes up a very heterogeneous landscape. To find out preparedness of partner organization to join common digital preservation system, we established ten different pilots of preparing and ingesting submission information packages with common specifications. These specifications define technical requirements for the digital objects, their metadata, and package structure to be submitted for digital preservation. In this paper, we show the piloting process, the experiences and the results, believing that these findings might be useful for various organizations involved with digital preservation.

Categories and Subject Descriptors
H 3.7 [**Information Storage and Retrieval**]: Digital Libraries – *Standards.*

General Terms
Experimentation, Standardization.

Keywords
Digital preservation, designated community, submission information package, metadata.

1. INTRODUCTION
The National Digital Library (NDL) of Finland [1, 2] is an entity within the remit of the Ministry of Education and Culture within the Finnish Government, which basis is formed by national libraries, archives and museums (partner organizations). Almost all memory organizations under the Ministry of Education and Culture of Finland are under an obligation to preserve the cultural material in their possession, of which a lot of is in a digital form. Most of this 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. The aim of the NDL is to ensure the effective and high-quality management, dissemination, and a common digital preservation (DP) 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 digital preservation system, promoting cooperation between the partner organizations, and building better services with open cooperation and expansion to include a large range of content. Common infrastructure and services will draw the practices of memory organizations closer, reduce the costs and fragmented nature of the systems, and intensify cooperation.

We are currently designing and implementing the NDL's DP system, which will be based on the OAIS reference model [3]. 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 the end of 2013 of the bit preservation will ensure that the digital materials in the possession of the partner organizations can be reliably preserved until the DP system becomes fully operational in 2016. In the second phase, the DP system 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 DP system for the research data.

NDL has defined specifications for preparing and creating unified Submission Information Packages (SIPs), with, for example, a redefined METS [4] schema and a closed set of acceptable file formats.[1] As we are building common DP system for various memory organizations, the unified structure enables efficient administration of the information on the long term and also enables semantically commensurable information content. In the NDL METS, some originally optional elements and attributes are stated as mandatory or as recommended, or the use has been restricted. From the acceptable file formats, some are recommended formats, which are straightforwardly accepted for preservation, where others are acceptable for transfer, which are migrated to a recommended format before preservation. The file format selections are mainly based on [5].

The preparedness of the partner organizations and the functionality of the specifications needed to be tested in practice, and therefore, the preparation and creation of SIPs were piloted with the partner organizations. This gave a lot of information about the partner organizations needs and the requirements needed for creating digital data according to the specifications. We believe that these findings might be useful for the various organizations involved with digital preservation. In this paper, we present the experiences and overall results of these pilots. In Section 2, we explain the structure of the pilots and collect the pilot experiences of the partner organization. In Section 3, we give the results found in the analysis of the SIPs. Finally, we conclude these pilots in Section 4.

[1] Specifications are available at http://www.kdk.fi/en, but only in Finnish.

2. PILOTS

To understand the preparedness of the partner organizations for SIP preparation and ingestion, ten pilots were established. Eight of the pilots included documenting the findings and practical work of creating SIPs for our DP system, and two of the pilots were fully reporting exercises. The selected partners (three libraries, five archives and two museums) have been involved with designing the NDL specifications, so they already had some background information about the DP system. All the pilots varied depending on the organization and the selected material, but the basic structure of each pilot was the following: The pilot started with a meeting, where the contact persons, timetable, pilot material types, duties and restrictions were agreed. In the first task, the partner organization identified the mandatory (and depending on time resources, also recommended) metadata fields from their back-end systems, with using the NDL's specifications. The partner organizations were supposed to list all the flaws they found from their system or from the specification, and suggest necessary enhancements to be taken into account in the specifications. In the second task, the organizations collected the test material from their systems, create one or several SIPs according to the specifications and submit them to the ingest pilot implementation. The organizations also wrote a report about their SIP creation experiences, such as listing the easy and difficult tasks, the needed changes in organization's processes or systems, improvement suggestions to the NDL's specifications and so on. The third part of the pilot was a task of the DP system designers, where the ingestion process was tested and documented. This included documenting all the found errors in the submitted packages, but also all the flaws found in the ingestion process. The last task of the pilot was to exchange experiences between the partner organizations and DP system designers.

2.1 Experiences of the partners

The partner organizations found the pilots interesting and useful. They experienced that their knowledge regarding various essential standards increased significantly. Further, the pilot gave them a lot of practical and concrete experience about the information packaging for DP. Three partner organizations found the NDL's specifications and packaging guidelines somewhat easy, inspiring them to create an automated process and choose a heterogeneous set of test material. Three other partner organizations found specifications and the required processes more demanding, and they decided to make the packages by hand. Two partners could provide the packages directly from their current systems, but in these cases, various differences were found against NDL's specifications. Some of the organizations found flaws in their current processes, such as digitizing without creating checksums, inconsistency with the documentation of the digitizing chain, or even missing provenance metadata.

The major feedback from the partner organizations was that providing several different types of examples would have been helpful. Some parts in our specifications were still under construction, such as how to present the rights metadata and the provenance information in various cases, or should different identifier definitions be nationally unified somehow. Also, more instructions were needed for the technical metadata and in several details containing controlled vocabularies. These partner organizations' experiences gave a lot of feedback to be analyzed for the work of the DP system, and as a result of these pilots, the NDL's specifications has been updated.

3. VALIDATION OF PACKAGES

According to the NDL's specification, the partner organization submits one or more SIPs in a ZIP archive to the DP system. The ZIP format is used only for the transfer step, making it possible to transfer one or more SIPs at once. In the first phase of the ingestion, the ZIP archive is unzipped and the structure of it is inspected (see Figure 1). Each first-level directory in the ZIP file is a SIP, which requires a valid METS document and a digital signature at the root of each first-level directories. If needed, there may be subfolders for the digital objects, for example for different manifestos of a given digital object. With this structure, each SIP can be validated separately. In the second phase, the digital signature included in the SIP is validated. The purpose of the digital signature is to validate the origin and the integrity of the data. In our final DP system, the ingestion will be terminated, if the SIP structure is incorrect or if it includes an incorrect signature. In the pilots, the inspection was continued to get all possible information (e.g., errors) from the ingestion. In the next step of the ingestion, the METS document is parsed against our METS schema, including all other schemas used inside the METS document (e.g., PREMIS [6]). In the pilots, the mandatory and recommended metadata fields were also inspected by hand, either entirely or by inspecting a few random samples from the METS document. After this, the checksums of the objects are validated, by comparing the calculated checksums of the digital objects and the checksums given in the METS document. In this phase, it is also inspected that all reported objects exist in the SIP and that there are no loose objects, that is, files without a reference from the METS document. The next step is to validate the file formats to ensure, that each file is correctly formed. After this, an inspection report is created and submitted to the partner organization. In various phases in the pilot, the validation was done with a custom Python or Java implementation including several 3rd party open source software, such as OpenSSL [7] for the signature validation or JHOVE2 [8] for the file validation. In the pilot, the report was done by hand, but automated reporting methods will be implemented to the production system. In our DP system, the last ingestion step is to create the Archival Information Package (AIP) from the SIP, but in these pilots, all the received data was removed from the server.

The validation against the METS schema is not fully enough for the METS documents, and various solutions are currently being built for more complex issues, which can be solved with Schematron [9]. Also, if using JHOVE for XML validation by default, it downloads all the required schemas from the internet for the validation, and therefore the process takes a lot of time. To solve this issue, XML catalogs [10] are required, so that the schemas are loaded locally. Also, the first phase of the packaging is a problem for one partner, where only huge movie files are managed, since creating a ZIP archive takes too much time, and it

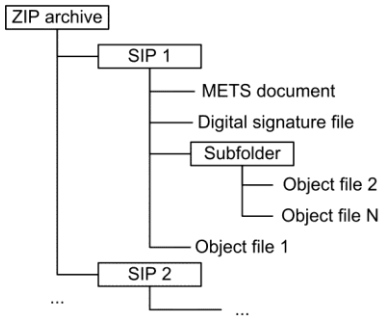

Figure 1. Structure of a ZIP including SIPs.

does not compress the already compressed data. However, this archiving step is used, so that the DP system knows, when the packages have been fully received. The only reasonable solution is to create an exceptional workaround with this partner.

3.1 Overall packaging results

In the analysis of the pilot, a grade between 0-2 was given for each SIP in each validation step as follows:

0. The part is missing or does not follow the specification,
1. The part includes severe errors or a large number of minor mistakes,
2. The part is flawless or includes only a few very minor mistakes.

Since the partner organizations submitted different number of SIPs, the average grade of each step was calculated for each organization separately. The average result of these organization grades for each ingestion step is shown in Figure 2.

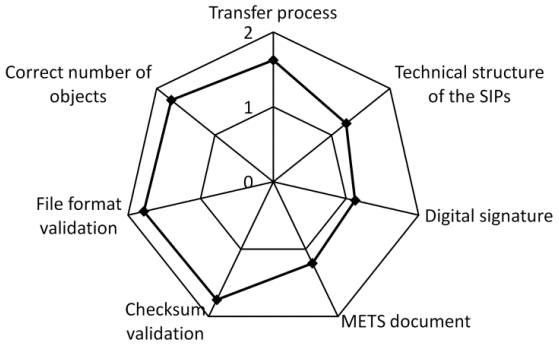

Figure 2. Average grades of the SIPs in validation steps.

From the Figure 2, it can be seen that usually in the SIPs the checksums were calculated correctly, the used file formats were correct and the SIP included the correct number of object files compared to the METS document references. However, the METS document creation had various types of small mistakes (see Section 3.2). The lack of examples generated uncertainty in the details. The creation of the METS document took a lot of time for some partner organizations. It was also found out, that in our specification the technical structure was a little bit confusing, and some of the organizations made various kinds of mistakes in this step, such as got confused, whether the ZIP archive is a SIP or a first-level directory inside a ZIP archive is a SIP.

3.2 Metadata results

In the NDL, a modified version of the METS schema is used, where more specific details have been added in the specification. In the NDL METS profile, the header, descriptive, technical, rights, provenance and struct map metadata are all mandatory, whereas the structural link and behavior metadata sections are forbidden. All the administrative metadata must be placed in a single administrative metadata section, so the use of several administrative metadata sections is denied. However, all the originally mandatory elements and attributes are still mandatory, and all elements and attributes are used in a way that it conforms to the original specification. The original idea was that when using the NDL METS schema, the resulted METS XML file is compatible to the official METS schema. However, as a result of the pilots, few additions were needed to the official METS specification. The request for these additions has been submitted to the METS board [11].

Figure 3 depicts how many organizations had different types of flaws related to the creation of the metadata. The most common mistake related to the creation of the METS document was one or a few missing or misused attributes (a). Five out of eight organizations had some flaws of this type. This is quite expected, since our METS profile includes a lot of mandatory or carefully defined attributes, and one or two of those might be forgotten or misunderstood in the first tryout. The external XML errors (b) were usually small, such as a single misused element or attribute. Namespace issues are quite difficult, and three organizations had problems in that part (cf. (c)). The specification of the provenance and rights metadata was partly under construction, and therefore some of the organizations did not pay attention on those parts (cf. (d) and (h)). Three of the organizations did not submit the digital signature (f), making it impossible to verify that the content of the METS document was correct. For clarity of Figure 3, let us mention that (a) or (n) are error types where one or a few incidental mandatory attributes or elements are missing. Even though for example error type (d) leads to missing attributes and elements, it does not affect to the count of error type (a) and (n), since it already is included in the error type (d).

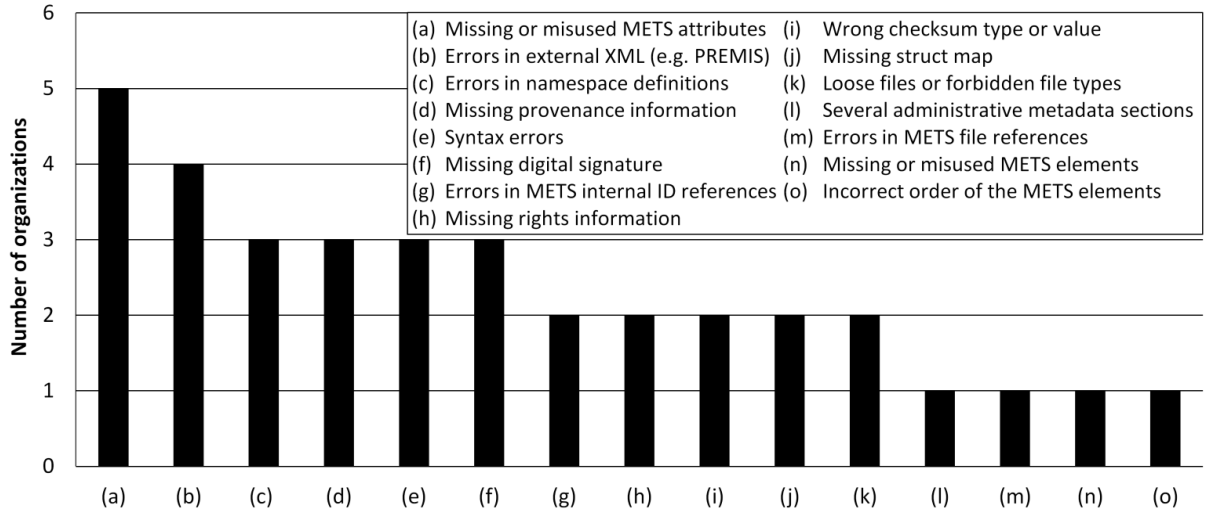

Figure 3. Flaws related to the received METS documents.

3.3 File format results

The file format validation is a process where it is examined whether the file is correctly formed or not. Our approach is to use existing 3rd party software for the validation, as far as possible. The validation was usually done for all the delivered files, but in few pilots, we needed to make the validation by random samples. The results are shown in Figure 4. From these file formats, PDF and MP3 are transferrable formats, where as all others are recommended formats. Most of the files were correctly formed. Some of the PDF/A files included a line feed character, which was a link directing to nowhere, and therefore the ingestion process discarded them. This raised a question, whether the errors of this type are acceptable or not, and how to deal with the issues of this type. We do not have a perfect solution for this, but the current plan is to decide and store a decision value for each error message, which defines a proper follow-up action. The forbidden file formats are file formats not allowed in the DP system, and therefore the ingestion process works correctly when defining those as faulty. The DP system must be built in a modular way, so that if better validation software is found or new file formats are accepted, the validation parts of these formats can be changed or added in a most convenient way.

As depicted in Figure 5, we received mostly JPEG-based files and XML-based files in the pilot. However, what is missing in the Figure 5 is how common certain file types are. Those partners who made the SIPs by hand provided only one or two files in their test packages, where as those partners who created or used an automated process, could provide more files. When JPEG2000 or XHTML files seemed to be quite popular based on Figure 5, but only one organization provided the files in those formats.

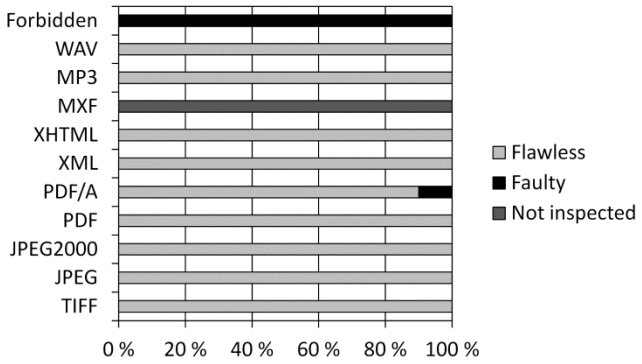

Figure 4. File format validation results.

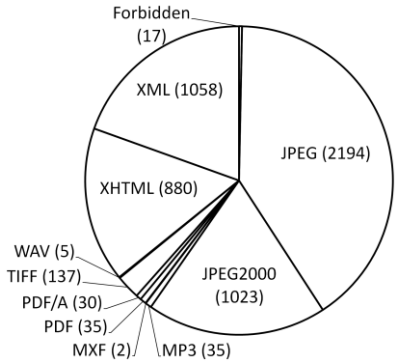

Figure 5. Received number of different file formats.

4. CONCLUSIONS

It was shown that the overall preparedness for preparing digital information for digital preservation is very different in different organizations. The object files are in good condition, but the flaws come up when packaging information and attaching necessary metadata to it. This practically means that the partner organizations are well prepared for a short-term usage of their digital data, but preparedness for a long-term DP needs development. However, it was shown that these problems are not overwhelmingly complicated, and with a carefully designed common technical support system, the partners are able to produce SIPs. One of the major focus point needed is the continuous updating and improvement of the NDL's specifications and operational methods, with paying attention to the required overall workload in package creation.

5. ACKNOWLEDGMENTS

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.

6. REFERENCES

[1] *National Digital Library*. URL=http://www.kdk.fi/en

[2] Helin, H., Koivunen, K., Lehtonen J. and Lehtonen K. 2012. Towards Preserving Cultural Heritage of Finland. In *Proceedings of the Cultural Heritage on line – Trusted Digital Repositories & Trusted Professionals* (Florence, Italy, December 10–14, 2012). NBN=http://nbn.depositolegale.it/urn:nbn:it:frd-9299

[3] *ISO 14721:2012: Open Archival Information System – Reference Model*. International Organization for Standardization. 2012.

[4] *METS Metadata Encoding and Transmission Standard*. URL=http://www.loc.gov/standards/mets/

[5] *Local Digital Format Registry (LDFR): File Format Guidelines for Preservation and Long-Term Access, Version 1.0*. Library and Archives Canada. 2010.

[6] *PREMIS Preservation Metadata Maintenance Activity*. URL=http://www.loc.gov/standards/premis/

[7] *OpenSSL – The Open Source Toolkit for SSL/TLS*. URL=http://www.openssl.org/

[8] *JHOVE2*. URL=http://www.jhove2.org/

[9] *ISO/IEC 19757-3:2006: Information technology – Document Schema Definition Language (DSDL) – Part 3: Rule-based validation – Schematron*. 2006.

[10] *XML Catalogs, OASIS Standard V 1.1*. Organization for the Advancement of Structured Information Standards. 2005.

[11] *METS Wiki - Change Requests.* URL=https://github.com/mets/wiki/blob/master/ChangeRequests/NDL-Finland-METS-changes.pdf