Modular Pre-Ingest Tool for Diverse Needs of Producers

Kuisma Lehtonen
CSC – IT Center for Science
P.O. Box 405, 02101 Espoo
Finland
Kuisma.Lehtonen@csc.fi

Pauliina Somerkoski
CSC – IT Center for Science
P.O. Box 405, 02101 Espoo
Finland
Pauliina.Somerkoski@csc.fi

Juha Törnroos
CSC – IT Center for Science
P.O. Box 405, 02101 Espoo
Finland
Juha.Tornroos@csc.fi

Mikko Vatanen
CSC – IT Center for Science
P.O. Box 405, 02101 Espoo
Finland
Mikko.Vatanen@csc.fi

Kimmo Koivunen
CSC – IT Center for Science
P.O. Box 405, 02101 Espoo
Finland
Kimmo.Koivunen@csc.fi

ABSTRACT

We introduce an open-source pre-ingest tool that assists the generation of Submission Information Packages (SIPs) that are to be submitted to the national digital preservation service in Finland. The pre-ingest tool consists of several independent components that produce the parts of a METS document required by the national preservation service. These components are easy to modify when developing services for different user demands or for different repositories. Users of the tool provide the necessary information as parameters for the tool, which produces the structure and descriptions for the SIP. The pre-ingest tool reduces the need to deeply understand either METS, PREMIS or other metadata formats to be able to preserve digital assets.

CCS CONCEPTS

• **Information systems** → Information systems applications → Digital libraries and archives • **Information systems** → Open source software

KEYWORDS

Digital Preservation Tools, Pre-Ingest, Open-Source Software

1 INTRODUCTION

Preparing and ingesting digital assets in an appropriate format to a preservation service can be a demanding task, especially in cases in which the producer is not familiar with the various preservation standards and metadata formats. This can be very time-consuming and therefore a very costly process. Thus, we have developed a pre-ingest tool to make it easier to create Submission Information Packages (SIPs) programmatically, which helps our partner organizations (libraries, archives and museums) to ingest digital assets to our national digital preservation service. Our pre-ingest tool decreases the burden of preserving data technically and thus releases time, which can then be used to produce digital assets instead of being wasting on considering how to preserve such assets.

Our national digital preservation repository provides services for preserving the cultural heritage and research data funded by the Ministry of Education and Culture of Finland. Currently, the preservation service is available for libraries, archives, museums and other organizations in Finland preserving cultural heritage by statutory obligation. We will extend our customer base soon with various organizations that have the need to preserve data, mainly nationally produced research data including publications and research methods. Given the diversity of the user needs, the digital assets to be preserved make up a very heterogeneous whole and simultaneously require various and flexible solutions.

Our national digital preservation service, based on the OAIS reference model [1], has been in production since 2015. Currently, we have more than one million Archival Information Packages (AIPs) in preservation, which amounts to more than 100 terabytes. We have defined common national preservation specifications [2], which describe in detail how digital assets should be prepared before ingesting them to the preservation service, including requirements for metadata and file formats. The design, implementation and development of the national digital repository are done in close collaboration with partner organizations. Thus, we have ensured that the user requirements are fulfilled by the pre-ingest tool.

To assist ingesting digital assets, our flexible pre-ingest tool can be used to generate SIPs. The tool produces a METS document containing all the necessary metadata conforming to our national preservation specifications. The tool includes creating descriptive and administrative sections for a METS document, creating a structural map, automatically extracting technical metadata from

files into the PREMIS metadata format, digitally signing the SIP and finally packaging the SIP as a TAR or ZIP package.

The rest of this paper is organized as follows: in Section 2, we give background information regarding the reasons why we developed this tool and how it is related to other pre-ingest tools. Section 3 presents the functionality of our tool. Finally, in Section 4 we conclude the paper and outline our future plans.

2 BACKGROUND

Ingesting digital assets into a preservation system can be a significant burden for some organizations, especially those with insufficiently competent IT staff. Therefore, we needed to simplify the pre-ingest process by providing a software tool for this. Earlier, in 2013, we found [3] that national memory organizations may have problems in creating syntactically and semantically correct SIPs with a valid METS document. Further, some organizations found the process of creating valid SIPs very time-consuming. The most common mistakes were related to the creation of a METS document with missing mandatory attributes or a document that misuses them. In addition, it was somewhat common to have errors in namespace definitions and internal references in the METS document. These somewhat trivial errors (for those having sufficient knowledge of the necessary metadata formats) are easy to prevent by using the pre-ingest tool to produce a valid METS document which contains most of the "low-level" tasks of creating a SIP. Based on these experiences, and with close collaboration with national memory organizations, we defined the functional and non-functional requirements for the pre-ingest tool.

There are various tools for creating SIPs for various repositories. For example, RODA-in [4] is an application for creating SIPs from local files and directories that includes batch-processing features. RODA-in provides an easy-to-use graphical user interface for creating SIPs but only supports Bagit and E-ARK SIP formats. The Rosetta SIP Factory [5] is a tool for creating SIPs suitable for the Rosetta preservation software [6]. We have chosen a more modular structure that constructs the METS document in parts, which gives more flexibility with which to integrate with back-end systems. The DURAARK WorkbenchUI [7] has been developed to assist the pre-ingest of architectural 3D data to the DURAARK system. It contains a graphical user interface and is implemented with Java. It supports BagIt and Rosetta SIP formats. In addition, some other preservation systems include functionalities for preparing the data for ingest, such as Archivematica [8], which supports the BagIt SIP format.

It is essential for us to avoid monolithic or complex workflows. In our needs, modularity, flexibility for modifications and the possibility to integrate and automatize the SIP creation process for partner organizations' back-end systems are the key issues. The diversity of back-end systems and processes, and the variety of metadata standards used in partner organizations lead to modular implementation of the pre-ingest tool.

As noted in [9], our digital preservation service is designed to receive large amounts of digital assets using a carefully designed and automated validation process at the ingest phase (see [10] for details). However, when the volume of data increases (both in the size and number of SIPs), it is crucial to be able to also automate the pre-ingest phase. Organizations that need to send a large amount of data may integrate their back-end systems with our national digital preservation system. Our pre-ingest tool can be deployed for this integration work very flexibly.

Our tool aims to create different pre-ingest services for our digital preservation service. Since the tool consists of several independent components, it is relatively easy to modify it or replace some components for different needs, including cases when developing services for different producer needs or for different repositories. Further, pre-ingest services with a GUI can be implemented for organizations that occasionally need to use the pre-ingest tool.

3 FUNCTIONALITY

Our pre-ingest tool is a set of modular software components. These components produce parts of a METS document and eventually produce a SIP that conforms to our national specifications. We have published the tool at GitHub as open-source software [11]. The architecture of the pre-ingest tool is selected to support easy service generation on top of it and to support the integration of a diversity of customer systems into the national digital preservation service while also allowing independent use when full system integration is too heavy. Pre-ingest services that are built on top of the pre-ingest tool employ the components of the tool in an appropriate order. Further, pre-ingest services may implement supplementary functionalities (e.g. sophisticated error handling and detailed reporting) whenever necessary.

An overview of our pre-ingest tool is depicted in Figure 1. To build a SIP conforming our digital preservation specifications, one can perform the following steps in the following order: *a*–*g* (where steps *a*–*d* can be executed in any particular order). As a result, these steps produce a digitally signed METS document, along with the digital assets composing a complete SIP that the producer can ingest. In what follows, we describe these steps in more detail.

Firstly, the tool generates descriptive metadata (step *a*). The descriptive metadata must be in a separate file and in a format acceptable in our repository (e.g. MARC, DC, MODS, EAD, EAC-CPF, LIDO, VRA or DDI). The location of metadata is given as an argument, and the tool encapsulates it into the METS descriptive metadata section. Descriptive metadata can describe a single file or a collection of files (e.g. a digitized book consisting a TIFF image per page).

Secondly, the tool generates technical metadata (*b*) for the files to be ingested. The technical metadata sections of the METS document are generated using an appropriate technical metadata format (e.g. PREMIS, MIX, VideoMD or AudioMD). The tool uses open-source components, for example the well-known JHove tool [12], to extract the necessary technical information from files. Should the technical metadata already be available in the organization's back-end system(s), the producer can easily modify the tool in order to retrieve this information from the back-end system automatically.

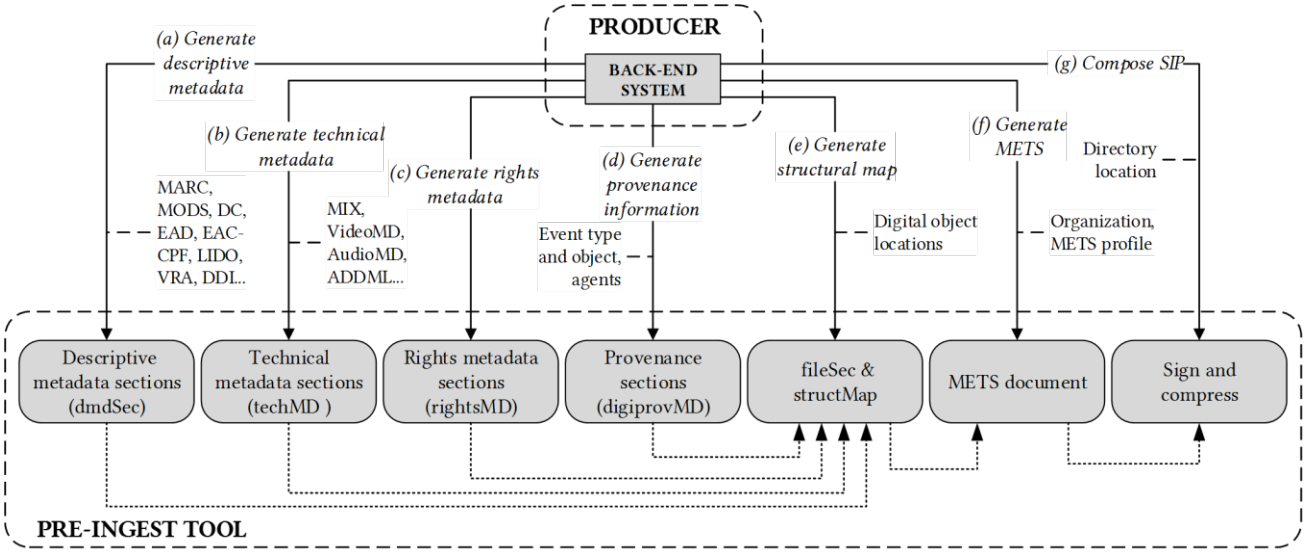

Figure 1: An overview of the pre-ingest tool

Thirdly, the tool generates rights metadata for digital assets (*c*). The required information is given as an argument. Rights metadata, like the descriptive metadata, is metadata that cannot be generated automatically in most cases but the information is crucial when disseminating digital assets years, or tens-of-years, later.

The fourth step is the generation of provenance information (*d*). The provenance section or sections of the METS document are created in PREMIS by giving the necessary information for the events and agents as arguments. PREMIS events and agents can describe the provenance information of a single file or a collection of files.

As noted above, these first four steps (*a-d*) can be executed in any order as they are independent from each other. The output of these steps, however, has to follow some general rules so that latter steps can produce a valid SIP.

The fifth step requires input from the first four steps in order to generate a structural map and file elements (*e*) for the METS document. In this step, the tool automatically creates references from the structural map to descriptive metadata and references from file elements (the digital assets to be preserved) to provenance information, technical metadata and rights metadata. Figure 2 depicts the element linking in a METS document in our national METS profile. The structural map is created based on different profiles. By default, the directory structure of the digital assets is described in the structural map. However, for example, some organizations create structural maps based on a structure described using the EAD[1] metadata format.

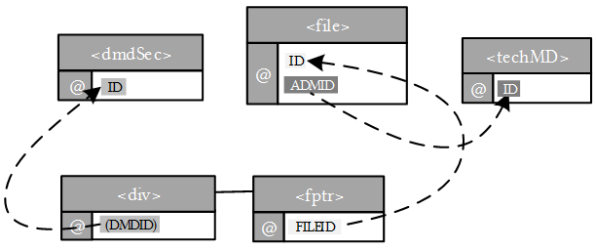

Figure 2: Element linkage in our national METS profile

Lastly, the tool generates the METS document (*f*) and composes the SIP (*g*). The tool collects the results of the previous steps and creates a complete METS document. The METS document is then digitally signed with a PKCS#7 signature[2] in order to ensure the integrity and fixity of the SIP during transfer into the preservation service. However, ingesting the SIP into the digital preservation service is not in the scope of the pre-ingest tool.

4 CONCLUSIONS AND FUTURE WORK

In Finland, the digital preservation of cultural heritage is enabled by a generalized preservation service for national libraries, archives and museums. The digital preservation service as an organization supports these memory organizations in several activities related to digital preservation. National memory organizations work in close collaboration with the preservation service in order to define national preservation specifications, which describe in detail how digital information shall be gathered in SIPs. In practice, this pre-ingest phase also requires several steps that are common for all organizations utilizing the preservation service. To help partner organizations in the pre-ingest phase, we have introduced a tool which simplifies the process of constructing SIPs. Especially, the pre-ingest tool significantly reduces the need for a deeper understanding of METS and PREMIS metadata formats and automates these common steps in the SIP's construction. With the

[1] Encoded Archival Description, https://www.loc.gov/ead

[2] https://tools.ietf.org/html/rfc2315

tool partner organizations can decrease the costs of the pre-ingest phase and therefore considerably lower the barriers to initiating the preservation of digital assets with the generalized preservation service.

We have created the pre-ingest tool in close collaboration with partner organizations. The requirements are diverse – one organization may even have multiple systems integrated into the national digital preservation service – so the pre-ingest tool must be modular enough to fulfil the different needs.

The pre-ingest tool is in production in partner organizations in order to integrate their back-end systems with the national digital preservation service. We have received good feedback when testing it with a representative sample of partner organizations. The pre-ingest tool will be used more widely when new partner organizations deploy our digital preservation service.

The pre-ingest tool will be developed further when national preservation specifications are updated. For example, support for new file formats, metadata standards and checksum algorithms can be implemented when needed. In the near future, there will be a need for more sophisticated SIP creation services. For example, the Open Science and Research Initiative[3] (a national initiative for promoting research information availability and open science) has plans to start the digital preservation of research data. The pre-ingest tool can be used as a basis for easy-to-use SIP creation services aimed at different disciplines.

ACKNOWLEDGEMENTS

The authors would like to thank all members of the Digital Preservation team at CSC – IT Center for Science Finland, as well as the National Digital Library support group for their valuable comments and input during the preparation of the national digital preservation service. Lastly, we thank the Ministry of Education and Culture of Finland for funding this project.

REFERENCES

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

[2] The National Digital Library. 2007. Digital Preservation Specifications. Ministry of Education and Culture of Finland. Retrieved May 31, 2017 from http://www.kdk.fi/en/digital-preservation/specifications

[3] Juha Lehtonen, Heikki Helin, Kimmo Koivunen, and Kuisma Lehtonen. 2013. On preparedness of Memory Organizations for Ingesting Data. In *Proceedings of the 10th International Conference on Preservation of Digital Objects (iPres 2013)*, José Borbinha, Michael Nelson, and Steve Knight (Eds.), Lisbon, Portugal, 276-279.

[4] Aadi Kaljuvee, Alex Thirifays, Angela Dappert, Björn Skog, Boris Domajnko, Jernej Križaj, Jože Škofljanec, Miguel Ferreira, Tarvo Kärberg, Thor Dekov Buur, Torben Lauritzen, Kuldar Aas, David Anderson, and Miguel Ferreira. 2017. *Deliverable D3.4: Records export, transfer and ingest recommendations and SIP Creation Tools*. European Archival Records and Knowledge Preservation (E-ARK), Revision 1.0.

[5] GitHub - Rosetta SIP Factory. 2017. National Library of New Zealand. Retrieved May 31, 2017 from https://github.com/NLNZDigitalPreservation/rosetta_sip_factory

[6] Rosetta: Digital Asset Management and Preservation. 2017. Ex Libris Ltd. Retrieved May 31, 2017 from http://www.exlibrisgroup.com/category/RosettaOverview

[7] DURAARK – Durable Architectural Knowledge. 2017. Retrieved May 31, 2017 from http://duraark.eu

[8] Peter van Garderen and Courtney C. Mumma. 2013. Realizing the Archivematica vision: delivering a comprehensive and free OAIS implementation. In *Proceedings of the 10th International Conference on Preservation of Digital Objects (iPres 2013)*, José Borbinha, Michael Nelson, and Steve Knight (Eds.), Lisbon, Portugal, 84-87. http://purl.pt/24107

[9] Juha Lehtonen, Heikki Helin, Kimmo Koivunen, Kuisma Lehtonen, and Mikko Tiainen. A National Preservation Solution for Cultural Heritage. 2015. In *Proceedings of the 12th International Conference on Digital Preservation (iPres 2015)*, Christopher A. Lee, Jonathan Crabtree, Leo Konstantelos, Nancy McGovern, Yukio Maeda, Maureen Pennock, Helen Tibbo, Kam Woods, and Eld Zierau (Eds.), Chapel Hill, USA, 247-248. Handle: 11353/10.429524

[10] Juha Lehtonen, Kuisma Lehtonen, Aarno Tenhunen, Juha Törnroos, Ville-Pekka Vainio, and Mikko Vatanen. 2016. Impressed by Ingest – Efficient and Reliable Workflows. In *Open Repositories 2016 Conference*, Dublin, Ireland.

[11] GitHub – dpres-siptools. 2017. CSC – IT Center for Science Ltd. Retrieved May 31, 2017 from https://github.com/Digital-Preservation-Finland/dpres-siptools

[12] JHOVE. 2017. Open Preservation Foundation. Retrieved May 31, 2017 from http://openpreservation.org/technology/products/jhove/

[3] http://openscience.fi/