Towards Automated Pre-Ingest Workflow for Bridging Information Systems and Digital Preservation Services

[In Records Management Journal, Vol 29, 3, 289-304]

Parvaneh Westerlund, Ingemar Andersson, Tero Päivärinta*, Jörgen Nilsson
Information Systems, Luleå University of Technology, Sweden
*) also M3S Research Group, University of Oulu, Finland

Abstract

Purpose – This research aims at automating pre-ingest workflow for preserving digital content, such as records, through middleware that integrates potentially many information systems with potentially several alternative digital preservation services.

Approach – Our design research approach resulted in a design for a model and component-based software for such workflow. A proof-of-concept prototype was implemented and demonstrated in context of a European research project, ForgetIT.

Findings – The study identifies design issues of automated pre-ingest for digital preservation while using middleware as a design choice for this purpose. The resulting model and solution suggests functionalities and interaction patterns based on open interface protocols between the source systems of digital content, middleware, and digital preservation services. The resulting workflow automates the tasks of fetching digital objects from the source system with metadata extraction, preservation preparation, and transfer to a selected preservation service. The proof-of-concept verified that the suggested model for pre-ingest workflow and the suggested component architecture was technologically implementable. Future research and development needs to include new solutions to support context-aware preservation management with increased support for configuring submission agreements as a basis for dynamic automation of pre-ingest and more automated error handling.

Originality/Value – The paper addresses design issues for middleware as a design choice to support automated pre-ingest in digital preservation. The suggested middleware architecture supports many-to-many relationships between the source information systems and digital preservation services through open interface protocols, thus enabling dynamic digital preservation solutions for records management.

Keywords Long-term digital preservation, pre-ingest, automation, middleware, workflow

Paper type Technical paper.

Introduction

Pre-Ingest is the preparatory stage for transferring digital records from information systems to one or more long-term digital preservation systems (DPS) (Kärberg, 2015). During this stage, contents are prepared to comply with the requirements of the ingest function of an archival system (CCSDS, 2012) that receives the preserved content in a DPS. The pre-ingest phase is crucial because it affects all subsequent preservation activities (ibid.). However, preparing digital content for submission to a long-term preservation repository requires both time and effort. Content producers are often reluctant to make investments to meet detailed preservation and submission guidelines, while incomplete information and insufficient metadata documentation are causing excessive costs on the side of archives (Rosenthal et al. 2005). A manual approach to pre-ingest is not a suitable strategy for preservation of digital material in the long term (Ross 2012). Producers of digital records and preservation organizations need to co-operate for long-term digital preservation (DP), aided by tools that automatically capture metadata and support the appraisal process (Hedstrom and Jinfang, 2008). While the literature addressed the issue of pre-ingest automation a while ago, projects to develop pre-ingest tools and elements of varyingly automated solutions started to emerge not before the mid-2010s (Päivärinta et al.. 2015; Kärberg, 2016; Lehtonen et al., 2017).

Lehtonen et al. (2017), in context of establishing a national digital preservation service in Finland, addressed the need for developing modular, flexibly modifiable, and easy-to-integrate pre-ingest workflows, to receive digital content from several producer organizations and their potentially many information systems (ISs). Moreover, the employed DPSs will change over time as well (Afrasiabi et al., 2014). That is, a DP solution should help to configure, automate, and manage digital preservation workflows in a context of potentially many-to-many integrations needed to bridge ISs producing digital content and long-term DPSs. Päivärinta et al. (2015) suggested an overall conceptual model for a supporting middleware for such a context and denoted the design problem of supporting automation of information transfer from ISs to DPSs and vice versa. The objective of this research is to develop and demonstrate a workflow in such middleware for *automating pre-ingest* tasks in the context of (potentially) many-to-many interactions between ISs and DPSs. This paper presents the designed three main functionalities of the suggested pre-ingest solution: selecting digital objects from a source system in harmony with automated fetching mechanisms for preserved materials, preservation preparation with automated metadata extraction and creation and transfer of submission information packages (SIPs; for standard definitions according to terminology of the Open Archival Information System (OAIS) model, see CCSDS, 2012). The future challenges for research and development of automated pre-ingest solutions are addressed based on the experimental tests conducted in context of the ForgetIT project (Gallo et al., 2018). The key contributions include a proof-of-concept for a solution for the above-mentioned three functionalities with identified design challenges, software components, and their interaction patterns.

In the remainder of the paper, we first outline the related work on pre-ingest solutions after which we introduce the major design issues and challenges given in the project that formed the basis for our work. Thereafter, the paper reports design of the pre-ingest architecture and workflow describing the interaction patterns between its software components with the results of preliminary tests. The paper rounds up with a discussion on the contributions and future research.

Related work

With an increasing amount, size and complexity of digital content, it is not feasible to manually deal with the preparation of material for digital preservation, considering the cost of staff as well as the complexity of manual processes (Ross, 2012; Kärberg, 2016; Lehtonen et al., 2017). Therefore, there is a need for automating the pre-ingest processes while involving archiving staff only when a human decision is necessary (Hedges et al., 2009).

As a fundament for developing solutions for preserving digital content transferred from other information systems, standardization efforts for metadata and interchange of digital formats have enabled the development of interoperable DP solutions. The producer-archive interface specification (PAIS 2006) is a standard for formally defining the process of transferring digital information objects between data producer and archive. PAIS is a concrete implementation of the main part of the formal definition phase and the transfer phase defined in PAIMAS (Huc et al., 2004). The main contribution of PAIS is a definition of an abstract Submission Information Package (SIP) but it still requires specific mapping to metadata standards used by the receiving archives, such as PREMIS (PREMIS working group, 2005), MODS (Guenther, 2003), and METS (McDonough, 2006).

Several works suggest the need for automating the early stages of the preservation process. A number of early efforts aimed to support extraction of metadata (e.g. Ross and Hedstrom, 2005, Greenberg et al., 2005, Ross and Kim 2005) whereas typically neglecting other tasks of pre-ingest workflows (Kärberg, 2016). The Chronopolis project (Hutt et al., 2008) represents an early effort for developing preservation metadata in a grid-based preservation system. The project suggested a workflow model for determining the materials to deposit, agreeing on the submission format, and establishing source and transfer mechanisms. The Chronopolis pre-ingest workflow allows the submitted data to be non-compliant with the standardized Submission Information Package (SIP) definitions (Hutt et al., 2008). The PROTAGE (Preservation Organizations using Tools in

Agent Environments) project identified issues in the pre-ingest phase among agencies and archives to facilitate their work, automated metadata creation and extraction, use of automated tools and standards, automation in creation of submission packages, and support in the appraisal and transfer processes (Rosa et al. 2009). The idea of PROTAGE was to link digital objects to long-term digital preservation processes by using agent-based software technology (Hägerfors et al.,2009). The PROTAGE agent prototype did not gain momentum for further development due to its technical complexity (Kärberg, 2016).

The Estonian National Archives created a pre-ingest tool, the universal archiving module (UAM), that allows archivists in government agencies to prepare digital records for archiving. The UAM resides in the archivist's computer and supports format identification and characterisation, automatic generation of file level metadata, and migration (Kärberg, 2016). CAST is another pre-ingest tool designed to collect websites and support a semi-automated delivery process of submission packages between a producer and an archive (Andersson et al., 2011). Recently, Lehtonen et al. (2017) described a pre-ingest tool aiming at modular and flexible workflow configuration in context of delivering digital content to a Finnish national solution for digital preservation. Both Kärberg (2016) and Lehtonen et al. (2017) address emergence of tools for creating SIPs for different kinds of long-term preservation repositories since mid-2010s. The examples include such software as RODA-in (Kaljuvee et al., 2017), Rosetta SIP Factory (NLZ, 2017), and the DURAARK Workbench UI (DURAARK, 2017). However, the workflows of the previous solutions have been regarded as either monolithic or narrow, i.e., not easily configurable for changing source system contexts or target services, or complex (Kärberg, 2016; Lehtonen et al., 2017). The Estonian UAM tool (Kärberg, 2015) needs to be installed on the computer of the agencies delivering records to the national archive, while the Finnish solution (Lehtonen et al., 2017) takes good steps towards configuring pre-ingest workflows to be integrated between the Finnish national archive and many source organizations and systems with a modular solution for metadata extraction.

Additional implementation and evaluation efforts are needed to improve functionality and automation level of pre-ingest methods and tools for long term preservation of records beyond the solutions of the national archives – to cover parts of the pre-ingest workflow beyond the task of metadata extraction and to be possible to integrate dynamically also with potentially varying service providers of digital preservation.

Design

Design Challenges

This paper relates to the ForgetIT project that focused on digital preservation issues of organizational and personal knowledge and suggested mechanisms for managing forgetting and contextualized remembering of digital information (Kanhabua et al., 2013; Niederee et al., 2015; Gallo et al., 2018). The project involved both academic and industrial European partners, with the budget of ca. 9 million Euros in 2013 – 2016. (For more detail, see https://www.forgetit-project.eu/). The task of our research group, as a part of the project, was to develop support for smooth transition of digital content from ISs to DPSs (Andersson et al. 2015) under the overall object of synergetic preservation that aimed at development of new solutions for "smooth bi-directional transitions" of knowledge between active use of information and management of digital preservation. However, due to the inherent long-term perspective of preservation related solutions, the aim was not to build a strongly integrated, monolithic system for transition workflows, but rather a middleware solution based on dynamically interacting components on which pre-ingest workflow can be configured.

This development took place simultaneously with, but independently from, the work by Lehtonen et al. (2017), who also address the importance of modularity and component-based software solutions for pre-ingest automation. The well-known benefits of component-based software design include the increased reuse of software at the component level, more flexible further development and maintenance of the component-based system, decreased production cost, and shorter implementation cycles (e.g., Lau and Wang, 2007). The middleware-orientation in our design is needed to establish a hub between potentially many source systems with potentially more than one DP services (Afrasiabi et al., 2014; Päivärinta et al., 2015). Such design is needed to avoid the "spaghetti" structure (Smith and McKeen, 2002) that would result from point-to-point, rigid integrations directly between many source systems and DPSs. That is, we reasoned for developing a messaging middleware solution that would allow systems to interface to the integration broker, each, which would also perform such additional functionalities as metadata packaging, messaging logic and management of the transactions among the integrated systems and DP services (cf. Smith and McKeen, 2002; Lam, 2005).

One challenge in automated pre-ingest is to decide which digital objects are candidates for preservation. Not every object in a collection needs preservation, such as outdated documents, duplicated images, or images of low quality. Another approach is to do an evaluation based on the use of a digital resource. If a digital resource has remained unused for a specified time, it might be a candidate for long-term preservation. However, other factors such as assessment of the topicality of a digital resource, computation of its usage patterns, or other digital objects related to it, could also influence that decision (Kanhabua et al., 2013). Thus, to detect and distinguish important from unimportant objects, textual and visual analysis techniques are needed (Mezaris et al., 2014). A key design challenge from our viewpoint was to make the pre-ingest workflow able to interact dynamically with software components (developed by other partners of the ForgetIT project) monitoring the transition of preservation value of potentially targeted information objects.

Another major issue is to determine the scope and specification of metadata added to an archival unit. Contextual metadata could be both manually added at the IS side or automatically extracted from digital objects. Context information can be described in a variety of dimensions such as who, when, where or in other dimensions such as topic, entity space (persons, organizations, events), and document space (related objects) (Ceroni et al., 2014).

Other design issues and challenges included the use of metadata standards, use of established transfer protocols, monitoring the use of DP services, and developing a strategy that manages changes in the middleware and in DPS over time (Afrasiabi et al., 2014).

While the overall reference model of the ForgetIT results is described in Gallo et al. (2018), this paper focuses on issue of pre-ingest automation in particular, taking the above-mentioned pre-requisites into account. A summarized list of initial design challenges that need to be addressed by the middleware thus included (Andersson et al., 2015):

A. Automated selection and fetching of candidate content for preservation based on pre-defined submission agreement, in interaction with software components supporting evaluation of preservation value.

B. Defining the scope, data, and structure of content allowed in digital object.

C. Automation of contextual metadata extraction for Submission Information Packages (SIPs).

D. Applying open interface protocols to support communication and transmission of content between systems deployed on various technological platforms.

Pre-Ingest Architecture

In order to address previously identified challenges where there is a need for a broker (middleware) between potentially many information systems (IS) and many digital preservation systems (DPS), the pre-ingest software component architecture is made up of three sections: *Information Systems*, *Middleware* and *Digital Preservation System*. To get an overview which software components resides where, a component architecture diagram was created (Figure 1). Most software components belong in the middleware but also needs to interact with components in IS and DPS. A short description of each software component in Figure 1 follows:

The **IS Adapter** and **Ingest** are interfaces between the two systems interacting with each other. The communication is handled technically with Content Management Interoperability Services (CMIS) (Brown, 2010) and REST (Fredrich, 2010).

The **Enterprise Service Bus (ESB)** supports communication between middleware components as well as with information systems (IS) and digital preservation system (DPS). The communication is handled by messages sent to and from message queues assigned to specific tasks.

The **Collector/Archiver** supports fetching digital content from IS to middleware. This component and the IS need to agree upon an appropriate transfer adapter. The Collector/Archiver is also responsible for assembling every object and metadata needed for creating and transferring submission information packages (SIP) to DPS (Andersson et al., 2014).

The **Extractor** retrieves information from different digital sources as input (e.g. text, image, or collections of them) and provides output as text or XML. The functionalities of extractor include entity extraction from text, concept detection in images, and visual quality assessment of images.

The **Condensator** takes in the output from the extractor and original objects to perform further linguistic text and image analysis, face detection and clustering. Its output is condensed analysis result in text, XML or image files (Mezaris et al., 2014).

The **Contextualizer** takes in the output from the extractor as well as original objects for utilisation of sufficient context metadata. If necessary, it makes use of external data sources on the web for enriching the context metadata. Its output is XML encoded data (Ceroni et al., 2014).

The **Metadata Repository** is a database management system that stores metadata for individual objects or collections and makes them available to other middleware components.

The **Staging server** is a dedicated physical space on a server that keeps digital objects which are managed during the middleware process.

The **Preservation DataStores** (PDS) by utilizing generic cloud storage, prepares content that is to be stored by different storage providers. The PDS includes a storlet engine that can be plugged into various cloud storages to perform format transformations, redundancy detection, aggregation processes, and integrity checks (Rabinovici-Cohen et al., 2013).

In addition to the components that relate directly to our middleware concept (Figure 1), other meaningful components need to be flexibly added to the pre-ingest solution architecture. As an example the Forgettor component (Niederee et al., 2015; Gallo et al., 2018) was implemented in the project to assist in the appraisal process by assessment of short- and long-term value of information resources.

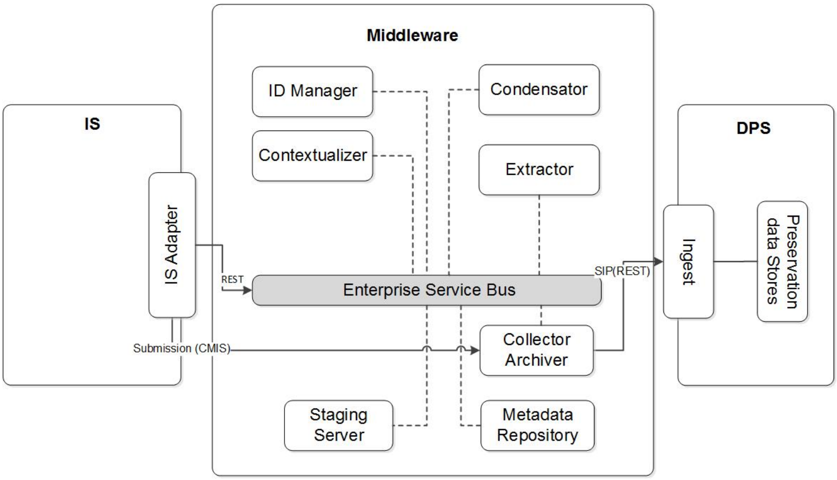


Figure 1 Pre-Ingest – Software Component Architecture

Pre-Ingest Workflow

Figure 2 presents a graphical representation of the activities within the pre-ingest workflow process. The diagram is used to get an overview over primary activities and relation between them. The workflow consists of three main steps (encircled with dashed lines) labelled in the figure as: 1) Selecting Objects, 2) Fetching & Preparation, and 3) Transfer & Ingest. Each vertical swim lane represents a participating system. Activities are drawn as rounded rectangles, a rectangle represents message queues that holds messages until they can be processed, and a diamond represents decision.

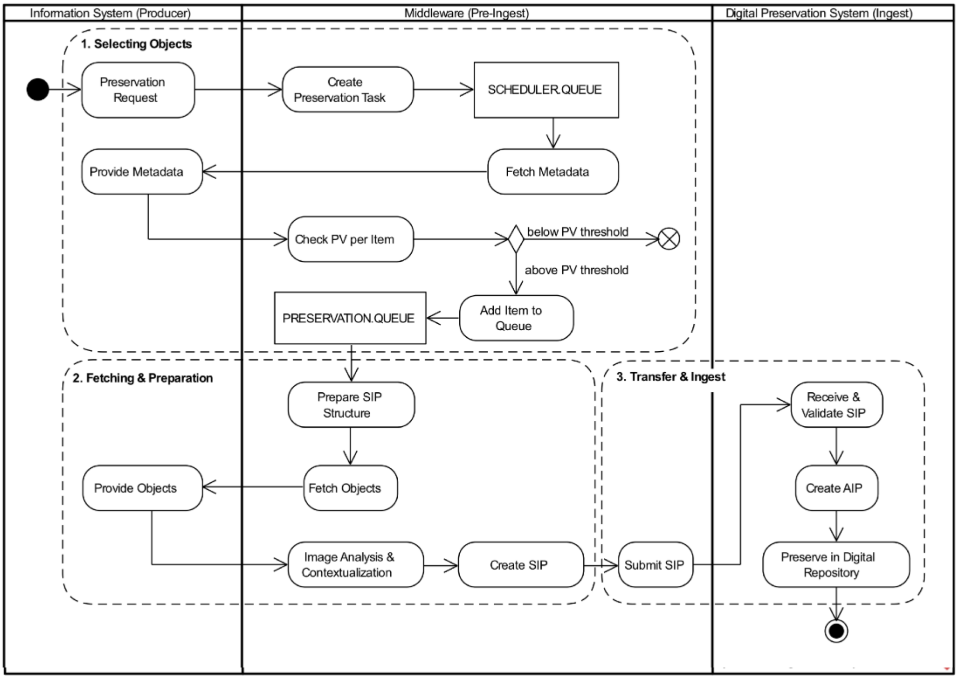


Figure 2 Activity Diagram for the Pre-Ingest Workflow

1. **Selecting Objects**: the workflow starts with a preservation request from the IS. With the assumption that a submission agreement for digital preservation is already established, including information on storage volume paid for, metadata requirements, package structure etc., a Preservation Task is created by the Enterprise Service Bus (ESB). This task is put in the Scheduler.Queue which the Collector/Archiver listens to. The Collector/Archiver reads the task from the queue and initiates fetching of metadata for the items. An Item is a generic descriptor of an object in the IS-Adapter and can thereby point to different kind of objects. When metadata from the IS-Adapter have been fetched, the selection of content for preservation commences. Whether an item is selected and placed in the Preservation.Queue depends on thresholds stated in the submission agreement, indicated by two measurements. The measurements are based on the concepts suggested by Niederee et al. (2015). Short-term value, labelled as Memory Buoyancy (MB) (Niederee et al., 2015), adapts to changing needs and interests, considering usage patterns, and information decay. Preservation value (PV) (Niederee et al., 2015) is a computed value based on several factors such as usage frequency, age of object, and related objects etc. PV is used to decide, how much to invest into the preservation of a resource or whether to preserve at all, thereby most relevant value for this part of the process (for more detail, see Kanhabua et al., 2013; Niederee et al., 2015). At this stage, objects could also be filtered out for other reasons stated in the submission agreement (e.g. according to acceptable file formats).
2. **Fetching & Preparation**: the Collector/Archiver reads from the Preservation.Queue and prepares a submission information package (SIP) folder structure according to the submission agreement. The Collector/Archiver retrieves the objects from IS-Adapter and stores them on the Staging Server, extracts metadata and stores it in the Metadata Repository. Later, external contextual metadata can be added to the objects in the Image Analysis & Contextualization activity, handled by the Extractor, Condensator, and the Contextualizer components. The submission agreement specifies, e.g., which image analysis method should be used, thresholds for image clustering, and which metadata specifications to apply (e.g. as a combination of METS,

MODS and PREMIS). The Collector/Archiver finalizes this stage with the creation of a SIP.

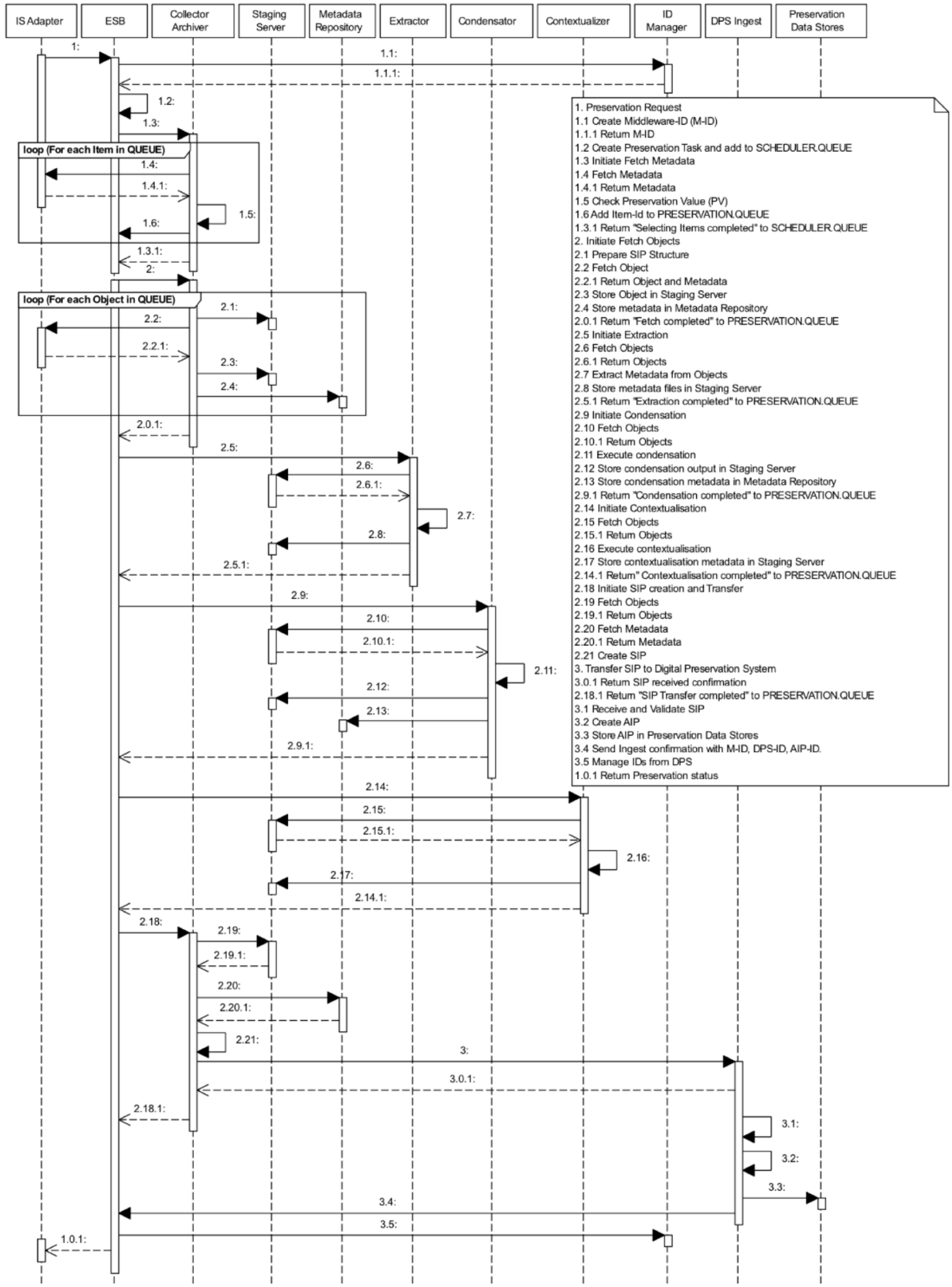
3. **Transfer & Ingest**: The submission agreement holds information on the service endpoint for the Digital Preservation System (DPS) and the Collector/Archiver uses this information to submit the package to the ingest of DPS. The SIP is validated upon reception in the DPS and then an Archival Information Package (AIP) is created and stored in Preservation Data Store. When all is done, a receipt is sent back to the Producer (IS-Adapter).

Source code for the prototype is available at https://www.forgetit-project.eu/en/project-results/code/ (accessed 2018-04-25).

Pre-Ingest Component Interactions

Figure 3 shows a sequence diagram. A sequence diagram shows software component interactions arranged in time sequence. It depicts the software components involved in the pre-ingest workflow scenario described in figure 2 and the sequence of messages exchanged between them to carry out the functionality needed. Every horizontal arrow in the figure represent a message being sent between components. These messages are numbered and briefly described in the textbox. Components are represented as vertical life-lines showing their duration during execution as a thin white rectangle on the life-line. An arrow pointing back to the same life-line represents an execution of functionality within the same component. The initial number of each message relates to the same main steps in figure 2 namely: 1) Selecting Objects, 2) Fetching & Preparation, and 3) Transfer & Ingest.

1. The pre-ingest process is initiated by a Preservation Request from the IS-Adapter to the Enterprise Service Bus (ESB). A preservation request could be initiated for different reasons: when content is no longer in use (low MB), upon creation for very valuable content, by scheduled preservation for all content above a predefined PV threshold, or by a manually triggered preservation request. Every preservation task needs a unique identifier (M-ID) provided by the ID-manager. This ID is used in every subsequent message to keep track of the process and to hold content together during pre-ingest. Since a preservation request can contain a collection of items, messages 1.4 to 1.6 iterates (loop) over all items in the collection, checking preservation value for each item and adds them to the preservation queue if the value is higher than the threshold. During this step there is also a check that the items are within the scope of expected submission, according to agreement. When all items have been processed the ESB is notified that selection of items is finished.
2. Having received the message that selection is finished, the ESB initiate fetching of objects by sending a message to the Collector/Archiver. The Collector/Archiver prepares a folder structure on the Staging Server for storage of objects and metadata. The Collector/Archiver then iterates (loop) over the list of objects and fetches them from the IS-Adapter. During fetching of objects, it also collects related metadata provided by the IS-Adapter. When all objects are fetched the ESB is notified and the process of metadata extraction and image analysis starts. This process has to be executed in order of Extraction, Condensation, and Contextualisation since they are dependent on the output from previous steps. When contextualisation is finished the ESB triggers the process of Submission Information Package (SIP) creation and transfer, with a message to the Collector/Archiver. The SIP is created with structure and metadata specifications in accordance to the submission agreement.
3. When Collector/Archiver finished the creation of SIP it transfers the package to the Ingest function of Digital Preservation System (DPS). After validation of the SIP an Archival Information Package (AIP) is created and stored. The identifier of the AIP, the identifier of the DPS, and the M-ID is returned back by a message to ESB to signal a successful ingest. These IDs are stored by the ID-manager to keep track of where objects from different ISs reside.



1. Preservation Request
1.1 Create Middleware-ID (M-ID)
1.1.1 Return M-ID
1.2 Create Preservation Task and add to SCHEDULER.QUEUE
1.3 Initiate Fetch Metadata
1.4 Fetch Metadata
1.4.1 Return Metadata
1.5 Check Preservation Value (PV)
1.6 Add Item-Id to PRESERVATION.QUEUE
1.3.1 Return "Selecting Items completed" to SCHEDULER.QUEUE
2. Initiate Fetch Objects
2.1 Prepare SIP Structure
2.2 Fetch Object
2.2.1 Return Object and Metadata
2.3 Store Object in Staging Server
2.4 Store metadata in Metadata Repository
2.0.1 Return "Fetch completed" to PRESERVATION.QUEUE
2.5 Initiate Extraction
2.6 Fetch Objects
2.6.1 Return Objects
2.7 Extract Metadata from Objects
2.8 Store metadata files in Staging Server
2.5.1 Return "Extraction completed" to PRESERVATION.QUEUE
2.9 Initiate Condensation
2.10 Fetch Objects
2.10.1 Return Objects
2.11 Execute condensation
2.12 Store condensation output in Staging Server
2.13 Store condensation metadata in Metadata Repository
2.9.1 Return "Condensation completed" to PRESERVATION.QUEUE
2.14 Initiate Contextualisation
2.15 Fetch Objects
2.15.1 Return Objects
2.16 Execute contextualisation
2.17 Store contextualisation metadata in Staging Server
2.14.1 Return" Contextualisation completed" to PRESERVATION.QUEUE
2.18 Initiate SIP creation and Transfer
2.19 Fetch Objects
2.19.1 Return Objects
2.20 Fetch Metadata
2.20.1 Return Metadata
2.21 Create SIP
3. Transfer SIP to Digital Preservation System
3.0.1 Return SIP received confirmation
2.18.1 Return "SIP Transfer completed" to PRESERVATION.QUEUE
3.1 Receive and Validate SIP
3.2 Create AIP
3.3 Store AIP in Preservation Data Stores
3.4 Send Ingest confirmation with M-ID, DPS-ID, AIP-ID.
3.5 Manage IDs from DPS
1.0.1 Return Preservation status

Figure 3 Sequence Diagram of Component Interactions

Testing and Identified Future Challenges

The scenario for the first test was to pre-ingest content from the PIMO (personal information management model; Maus and Schwarz, 2014) system to a DPS. The PIMO vendor was a partner of the ForgetIT project. PIMO supports management of photo collections enhanced with context information based on a semantic ontology (Maus and Schwarz, 2014). One of the benefits of the

PIMO approach is the integration with user data via dedicated applications or plug-ins for standard applications such as browsers, address books, calendars, email, or file system. The Semantic Desktop of PIMO allows extensive logs of user actions, including whether an object has been annotated or viewed, searches, archive access, and any external events that were found through integration with browser and email software (if the users enable this feature).

In an experiment prior to preservation actions, ten participants used PIMO user interface to Semantic Desktop to manage their photographs. They took 40–80 images, which were organised into specific events (as photo collections), to the first session. In the second session, participants performed a few tasks on their two largest photo collections. The tasks included: 1. to review and change preservation preferences, 2. to apply one of the detected visual concepts as a filter, 3. to add a note describing the collection as a whole with key concepts, 4. to annotate individual photographs using concepts to make them more searchable, and finally 5. to search PIMO for a user-defined concept. In the third and final session, participants reviewed the time capsule generated by the Semantic Desktop based on evidences gathered in the first two sessions.

Analysing the data gathered during the experiment, a conclusion was made for imperative steps required for content transformation to a DPS (in the presence of Semantic Desktop case of this study). These steps were addressed and demonstrated by the above-mentioned functionalities of the pre-ingest workflow in the middleware. These results relate to the four design challenges previously expressed in this paper as follows:

- The middleware starts by inspecting settings in the submission agreement, with PV as the first factor to consider. In this case, PV oriented in the PIMO experiment. This process addresses the first design challenge (A) as the middleware could work on the value for the PV metric based on which a content is determined as a candidate for preservation.
- Next, the middleware uses the submission agreement to arrange which components will be involved in the pre-ingest workflow for this content and the workflow's route within the middleware. This step addresses the design challenge B.
- The middleware preserves content automatically according to the submission agreement. This step covers challenge C in the list of design challenges by automatically performing the tasks required for creating a SIP and including extraction of context metadata.
- The middleware components interact with each other, other systems, and external services using appropriate interfaces for metadata extraction. The final challenge (D) is addressed by this functionality which makes such interactions possible in the workflow.
- The design allows for revisiting and adapting preservation strategy to check if enhancements of Semantic Desktop-enabled applications or plug-ins are installed by the user which leads to new rules in the submission agreement.

The tests in general confirmed the desired functionality of a workflow where a collection of selected images was automatically fetched from PIMO, according to the submission agreement and interacting with the Forgettor component functionalities, through the middleware to a receiving DPS. The test demonstrated an automated process where context metadata was extracted from the objects as well as from external sources (such as date and time of the photos, location, etc.). Regarding the automatically generated visual concepts, participants generally found that they were in line with their own judgment at detecting aspects of their images. A SIP with structure and content was generated and transferred to receiving DPS without technical interruptions.

The experimental tests identified also a number of additional design challenges to be addressed by further research and development initiatives:

I. *Adaptation on the IS side to support the use of selected interoperability interface.* DP services are presented through interfaces to ISs. Interfaces have an important role especially in many-to-many interactions between ISs and DPSs. Therefore, it is crucial for both ISs and DPSs to be able to adapt to varying interfaces that are offered by DP services. Indeed, it is of same importance for the interfaces to follow some kind of standard for the IS to have to adapt to the same configurations at different times.

II. *Determination of where to execute metadata extraction*; on the IS side, in middleware, or in a remote server as a service? In an IS, specific activities that are central to the ingest function can be handled. For example, in a content management system, creation of metadata can be provided with the content capture process and is maintained over the content's lifecycle (Korb and Strodl, 2010). Such metadata cannot be directly inserted into metadata section of a digital object since its format are often not compatible with digital preservation standards (Korb and Strodl, 2010), such as OAIS (CCSDS, 2012). According to additional requirements in submission agreement, more metadata might be extracted in the middleware, some of which might be extracted using remote DP services (e.g. in the cloud).

III. *Supporting alternative workflows such as error and rejection processes and routing customization of DP services*. This research aims at maximally automating many-to-many interactions between ISs and DPs. Even though we claim that this can be achieved through our design of workflow in the middleware, there are still considerations to be taken. Currently, there is a need for human interference in case of errors in the workflow, rejection of a digital object, or routing DP services. These tasks should be automated as well in the production solutions.

IV. *The need for applying security mechanisms at transfer aiming for authenticity protection of content on-hold during middleware processing*. DP mechanisms should establish the identity of content, services, and users interacting within the environment, in addition to manage intellectual property rights and privacy, and to secure the integrity and authenticity of content and services (Lavoie and Dempsey, 2004). Such concerns need to be involved in the processes in the middleware and remote DP services as well.

V. *Verifying that content* (expected file format, number of instances, size, etc.) *to be sent from IS to middleware is according to a submission agreement* before transfer takes place. A submission agreement has an essential role in making decisions regarding what processes are going to be executed on a digital object. Accordingly, it is of a great importance that the content of a digital object is in accordance specifications of the submission agreement assigned to the digital object.

Discussion and Concluding Remarks

This study introduced a model for workflow and components in the middleware required to automate pre-ingest tasks for transferring digital content from information systems (ISs) to digital preservation services (DPSs). Creating consistent information packages together with improvements in automation of pre-ingest workflow was demonstrated. The experimental tests of the middleware confirmed that implementation of the workflow was technologically possible and allowed us to uncover challenges and new opportunities that will contribute to the further development of pre-ingest middleware. Our component based prototype for the workflow solution focused especially on improving the functionality of

1) fetching digital content from the source IS, being able to interact with components determining the preservation values automatically and utilizing the pre-set submission agreement,
2) automating preservation preparation with automated metadata extraction together with adding external context metadata, and
3) transferring the resulting SIPs to DPSs.

Comparing our solution with the recent systems developed for national archives (Kärberg 2015; Lehtonen et al., 2017), we argue that middleware-based solutions could improve possibilities for pre-ingest automation especially in contexts where organizations need to preserve their long-term records more dynamically and eventually utilize a variety of DPS vendors and solutions. While Lehtonen et al. (2017) suggest that their solution is providing components also to the situation where full system integration between ISs and DPSs may become too heavy (as in the cases of national archives in general), our aim is exactly to support such deeper (but dynamic) integration to the extent possible. For example, a potential target organization for our solution could be a public sector agency (or private company) which needs to preserve records from tens of ISs, e.g., based on legal compliance requirements, potentially to more than one type of DPSs that could be either internal or even partially outsourced. Anyhow, our solution should be interesting to such contexts in which automation of pre-ingest based on submission agreement specifications in general would be meaningful.

Alike Lehtonen et al. (2017), we denote the importance of splitting the pre-ingest activities into well-defined components. This enables component reuse in alternative workflow configurations.

As well, our design employs interfaces that support open standardized communication protocols between internal middleware components and between ISs and DPSs.

While universal archiving module (UAM) (Kärberg, 2015) demonstrates the possibility for adopting semi-automated tools for pre-ingest on the archivist workstation, the few first transfers of each content type to the national archives need more manual work to make subsequent submission workflows efficient. This study suggests a solution more suited for contexts where long-term preservation needs to be integrated more with on-going records management. In such milieus, a submission agreement should automatize further configuration of workflows between ISs and DPSs through middleware, e.g., in organizational contexts where several systems need to be flexibly aligned to potentially more than one DPSs. In UAM, different types of metadata are extracted through one single process while metadata extraction in our workflow is distributed among a few components for each type which we estimate to become both quicker and scalable. Compared to Lehtonen et al. (2017), our solution covers also the tasks of fetching the digital objects while defining the preservation value, providing contextual external metadata by image analysis, clustering and contextualization (if so desired), and transferring the SIP to the selected DPS according to the submission agreement.

Using middleware to ease many-to-many integrations among systems is of course not a new approach to systems development in general (e.g., Linthicum, 2000). Nonetheless, our study contributes in particular to the field of digital preservation of records by using standard interfaces to comply with different formats and standards on both sides, demonstrating capability of many-to-many communications, representing a step towards configuring workflows based on submission agreement, and demonstrating how these tasks can be automatically performed. The variety of potential content types together with producers' requirements for transfer to the middleware may lead to more specific submission agreements that impose conditions not yet considered in this study. One factor jeopardizing automation of pre-ingest is occurrence of an error in the workflow demanding human intervention to adjust the process. To achieve higher level of automation, the process of handling errors in pre-ingest should thus become automatic as well.

Logically, our solution should be able to support configurations of alternative workflow paths and flexible adaptation of DP services for selection of combinations of services for specific circumstances. In the future solution, the submission agreement will be the key by which to configure the subsequent workflow according to the specifications. However, as the actual configuration and management tasks for submission agreements themselves are not yet included in the solution reported in this paper, more research and development efforts are needed to promote the level of automation of pre-ingest and specially to refine preservation administration tasks based on submission agreements to support and to flexibly configure other components in a pre-ingest workflow. That is, our first experiment revealed the need for developing a new component, Context-aware Preservation Manager that will support more dynamic submission agreements and workflow configurations. While such a component was included in the ForgetIT reference model (Gallo et al., 2018), the results from development and experimentation with that component are reported elsewhere (e.g., Westerlund et al., 2018).

Acknowledgments

This work was partially funded by the European Commission in the context of the FP7 ICT project ForgetIT (under grant no: 600826).

References

Afrasiabi Rad, P., Nilsson, J. and Päivärinta, T. (2014), "Administration of Digital Preservation Services in the Cloud Over Time: Design Issues and Challenges for Organizations", in Dr Barbara Endicott-Popovsky (Ed.), *Proceedings of the 2nd International Conference on Cloud Security Management*, presented at the 2nd International Conference on Cloud Security Management, Thomson Reuters ISI, Reading, UK.

Andersson, I., Afrasiabi Rad, P., Lindqvist, G., Nilsson, J., Päivärinta, T., Rabinovici-Cohen, S., Maus, H., et al. (2014), *D5.2: Workflow Model and Prototype for Transition between Active System and AIS - First Release*, ForgetIT, available at: https://www.forgetit-project.eu/fileadmin/fm-dam/deliverables/ForgetIT_WP5_D5.2.pdf.

Andersson, I., Lindbäck, L., Lindqvist, G., Nilsson, J. and Runardotter, M. (2011), “Web Archiving Using the Collaborative Archiving Services Testbed”, in Cunningham, P. and Cunningham, M. (Eds.), *EChallenges E2011 Conference Proceedings*, IIMC International Information Management Corporation, Florence, Italy.

Andersson, I., Nilsson, J., Lindqvist, G. and Westerlund, P. (2015), *D5.3: Workflow Model and Prototype for Transition between Active System and AIS – Second Release*, ForgetIT, available at: https://www.forgetit-project.eu/fileadmin/fm-dam/deliverables/ForgetIT_WP5_D5.3.pdf.

Brown, A. (2010), “Content Management Interoperability Services (CMIS) Version 1.0”, available at: http://docs.oasis-open.org/cmis/CMIS/v1.0/cmis-spec-v1.0.html (accessed 15 July 2014).

CCSDS. (2012), *Reference Model for an Open Archival Information System*, CCSDS Secretariat / Magenta Book., Consultative Committee for Space Data Systems, Recommendation for Space Data System Practices, Washington, USA, available at: https://public.ccsds.org/pubs/650x0m2.pdf (accessed 12 November 2018).

Ceroni, A., Greenwood, M., Mezaris, V., Niederée, C., Papadopoulou, O. and Solachidis, V. (2014), *D6.2: First Release of Tools for Contextualization*, ForgetIT, available at: https://www.forgetit-project.eu/fileadmin/fm-dam/deliverables/ForgetIT_WP6_D6.2.pdf.

DURAARK. (2017), “DURAARK – Durable Architectural Knowledge”, available at: http://duraark.eu/ (accessed 26 April 2018).

Fredrich, T. (2010), “What is REST?”, available at: http://www.restapitutorial.com/lessons/whatisrest.html (accessed 24 August 2013).

Gallo, F., Niederée, C. and Allasia, W. (2018), “Bridging Information Management and Preservation: A Reference Model”, in Mezaris, V., Niederée, C. and Loggie, R. (Eds.), *Personal Multimedia Preservation*, Springer, pp. 183–229.

Greenberg, J., Spurgin, K. and Crystal, A. (2005), *AMeGA (Automatic Metadata Generation Applications) Project*, University of North Carolina, Chapel Hill, USA, available at: http://ils.unc.edu/mrc/amega.html.

Guenther, R.S. (2003), “MODS: The Metadata Object Description Schema”, *Portal: Libraries and the Academy*, Vol. 3 No. 1, pp. 137–150.

Hägerfors, A., Quisbert, H. and Nilsson, J. (2009), “Agent technology supporting digital preservation”, *International Multi-Conference on Engineering and Technological Innovation*, presented at the International Multi-Conference on Engineering and Technological Innovation, Orlando, USA.

Hedges, M., Blanke, T. and Hasan, A. (2009), “Rule-based curation and preservation of data: A data grid approach using iRODS”, *Future Generation Computer Systems*, Vol. 25 No. 4, pp. 446–452.

Hedstrom, M., Jinfang, N. and Marz, K. (2008), “Incentives for Data Producers to Create ‘Archive-Ready’ Data: Implications for Archives and Records Management”, *Proceedings from the Society of American Archivists Research Forum*, Vol. 30, Society of American Archivists, San Francisco, USA.

Huc, C., Boucon, D., Sawyer, D.M. and Garrett, J.G. (2004), “The Producer-Archive Interface Methodology Abstract Standard (PAIMAS)”, *Space OPS 2004 Conference*, Aerospace Research Central, Montreal, Quebec, Canada, available at:https://doi.org/10.2514/6.2004-649-446.

Hutt, A., Westbrook, B., Kozbial, A., McDonald, R. and Sutton, D. (2008), “Developing preservation metadata for use in grid-based preservation systems”, *Proceedings of the Fifth International Conference on Preservation of Digital Objects (IPRES 2008)*, St Pancras, London, pp. 145–150.

ISO. (2006), *ISO 20652:2006(E) Producer–archive Interface —Methodology Abstract Standard*, No. CCSDS 651.0-B-1:2004, International Organization for Standardization - ISO, Switzerland.

ISO. (2016), *ISO 15489-1:2016 - Information and Documentation - Records Management - Part 1: General*, International Organization for Standardization - ISO.

Kaljuvee, A., Thirifays, A., Dappert, A., Skog, B., Domajnko, B., Križaj, J., Škofljanec, J., et al. (2017), *D3.4; Records Export, Transfer and Ingest Recommendations and SIP Creation Tools*, European Archival Records and Knowledge Preservation.

Kanhabua, N., Niederée, C., Loggie, R., Tran, T., Djafari-Naini, K., Maus, H. and Schwarz, S. (2013), *D3.1: Report on Foundations of Managed Forgetting*, ForgetIT, available at: https://www.forgetit-project.eu/fileadmin/fm-dam/deliverables/ForgetIT_WP3_D3.1.pdf.

Kanhabua, N., Niederée, C. and Siberski, W. (2013), "Towards Concise Preservation by Managed Forgetting: Research Issues and Case Study", *Proceedings of the 10th International Conference on Preservation of Digital Objects*, presented at the 10th International Conference on Preservation of Digital Objects, Lisbon, Portugal, pp. 252–257.

Kärberg, T. (2015), "Digital preservation of knowledge in the public sector: a pre-ingest tool", *Archival Science*, Vol. 15 No. 1, pp. 83–95.

Kärberg, T. (2016), *Digital Preservation of Knowledge – a Theoretical-Practical Research at the National Archives of Estonia*, Doctoral Dissertation, University of Tartu, Tartu, Estonia, available at: http://hdl.handle.net/10062/54614.

Korb, J. and Strodl, S. (2010), "Digital Preservation for Enterprise Content: a Gap-analysis between ECM and OAIS", *Proceedings of the 7th International Conference on Preservation of Digital Objects*, presented at the 7th International Conference on Preservation of Digital Objects, Vienna, Austria, pp. 221–229.

Lam, W. (2005), "Barriers to E-Government Integration", *Journal of Enterprise Information Management,* Vol. 18 No. 5, pp. 511–530.

Lau, K.-K. and Wang, Z. (2007), "Software Component Models", *IEEE Transactions on Software Engineering,* Vol. 33 No. 10, pp. 709–724.

Lavoie, B. and Dempsey, L. (2004), "Thirteen Ways of Looking at... Digital Preservation", *D-Lib Magazine*, Vol. 10 No. 7/8, available at: http://dlib.org/dlib/july04/lavoie/07lavoie.html.

Lehtonen, K., Somerkoski, P., Törnroos, J., Vatanen, M. and Koivunen, K. (2017), "Modular Pre-Ingest Tool for Diverse Needs of Producers", *Proceedings of the 14th International Conference on Digital Preservation*, presented at the 14th International Conference on Digital Preservation, Kyoto, Japan.

Linthicum, D.S. (2000), *Enterprise Application Integration*, 1st ed., Addison-Wesley Professional, Massachusetts.

Maus, H. and Schwarz, S. (2014), *D9.2: Personal Preservation Mockups (Annex with Mockups)*, ForgetIT, available at: https://www.forgetit-project.eu/fileadmin/fm-dam/deliverables/ForgetIT_WP9_D9.2.pdf.

McDonough, J.P. (2006), "METS: standardized encoding for digital library objects", *International Journal on Digital Libraries*, Vol. 6 No. 2, pp. 148–158.

MD5. (2013), "MD5 Hash Generator", *Dan's Tools*, available at: http://www.md5hashgenerator.com/ (accessed 14 August 2015).

Mezaris, V., Solachidis, V. and Papadopoulou, O. (2014), *D4.2: Information Analysis, Consolidation and Concentration Techniques, and Evaluation - First Release*, ForgetIT, available at: https://www.forgetit-project.eu/fileadmin/fm-dam/deliverables/ForgetIT_WP4_D4.2.pdf.

Niederee, C., Kanhabua, N., Gallo, F. and Logie, R.H. (2015), "Forgetful Digital Memory: Towards Brain-Inspired Long-Term Data and Information Management", *SIGMOD Record,* Vol. 44 No. 2, pp. 41–46.

NLZ. (2017), “GitHub - Rosetta SIP Factory”, *National Library of New Zealand*, available at: https://github.com/NLNZDigitalPreservation/rosetta_sip_factory (accessed 26 April 2018).

Päivärinta, T., Westerlund, P. and Nilsson, J. (2015), “Integrating Contemporary Content Management and Long-Term Digital Preservation: A Design Problem”, *Lecture Notes in Business Information Processing*, Vol. 223, Springer, Oulu, Finland, pp. 92–107.

PREMIS working group. (2015), *Data Dictionary for Preservation Metadata*, Library of Congress, Ohio, USA, p. 237.

Rabinovici-Cohen, S., Marberg, J., Nagin, K. and Pease, D. (2013), “PDS Cloud: Long term digital preservation in the cloud”, *IEEE International Conference on Cloud Engineering (IC2E 2014)*, IEEE, Redwood City, CA, USA, pp. 38–45.

Riksarkivet. (2011), *The E-Archive and e-Diarium Project, EARD*, Riksarkivet, Stockholm, Sweden, available at: https://riksarkivet.se/Media/pdf-filer/Projekt/eARD_informationstext_eng.pdf.

Rosenthal, D., Robertson, T., Lipkis, T., Reich, V. and Morabito, S. (2005), “Requirements for digital preservation systems: a bottom-up approach”, *D-Lib Magazine*, Vol. 11 No. 11.

Ross, S. (2012), “Digital Preservation, Archival Science and Methodological Foundations for Digital Libraries”, *New Review of Information Networking*, Vol. 17 No. 1, pp. 43–68.

Ross, S. and Hedstrom, M. (2005), “Preservation research and sustainable digital libraries”, *International Journal on Digital Libraries*, Vol. 5 No. 4, pp. 317–324.

Ross, S. and Kim, Y. (2005), “Digital Preservation Automated Ingest and Appraisal Metadata”, edited by Thanos, C.*DELOS Research Activities 2005*, pp. 64–77.

Smith, J.D.. and McKeen, H.A. (2002). "New Developments in Practice II: Enterprise Application Integration", *Communications of the AIS,* Vol. 8 Article 31, pp. 451–466.

Westerlund, P., Andersson, I., Päivärinta, T. (2018) "Towards Automated, Context-Aware Management of Preservation Submissions", 41st Information Systems Research Seminar (IRIS), 5th to 8th of August 2018, Aarhus, Denmark.