DISSERTATIONES
HISTORIAE
UNIVERSITATIS
TARTUENSIS
39

TARVO KÄRBERG

Digital Preservation of Knowledge – a theoretical-practical research at the National Archives of Estonia

DISSERTATIONES HISTORIAE UNIVERSITATIS TARTUENSIS
39

DISSERTATIONES HISTORIAE UNIVERSITATIS TARTUENSIS
39

TARVO KÄRBERG

Digital Preservation of Knowledge – a theoretical-practical research at the National Archives of Estonia

UNIVERSITY OF TARTU
Press

Institute of History and Archaeology, Faculty of Arts and Humanities, University of Tartu, Estonia

The dissertation is accepted for the commencement of the degree of Doctor of Philosophy (in History) on October 25, 2016 by the Council of the Institute of History and Archaeology, University of Tartu.

Supervisor: Prof. Prof. Aadu Must
Institute of History and Archaeology
University of Tartu

Opponent: Dr. Anssi Jääskeläinen
Mikkeli University of Applied Sciences
Finland

Commencement: January 5th 2017 at 16.15, UT Senate Hall,
Ülikooli st. 18–204, Tartu, Estonia

ISSN 1406-443X
ISBN 978-9949-77-290-2 (print)
ISBN 978-9949-77-291-9 (pdf)

University of Tartu Press
www.tyk.ee

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ORIGINAL PUBLICATIONS

I. **Tarvo Kärberg.** Digital Preservation of Knowledge in the Public Sector: A Pre-Ingest Tool.
The article (DOI 10.1007/s10502-013-9211-z) has been published in journal Archival Science (ISSN: 1389-0166).

II. **Tarvo Kärberg.** Digital Preservation and Knowledge in the Public Archives: For Whom?

The article (DOI 10.1080/23257962.2014.942606) has been published in journal Archives and Records: The Journal of the Archives and Records Association (ISSN 2325-7962 (Print), 2325-7989 (Online)).

III. **Tarvo Kärberg.** Toward Common Ontologies of Facets of the Archival Access Portal.

This article (DOI 10.1007/978-3-319-03437-9_36) has been published in Springer CCIS (Communications in Computer and Information Science) series (ISBN 978-3-319-03436-2 (Print), 978-3-319-03437-9 (Online)).

IV. **Tarvo Kärberg, Koit Saarevet.** Transforming user knowledge into archival knowledge.

The article (DOI: 10.1045/march2016-karberg) has been published in D-Lib Magazine (ISSN: 1082-9873).

This doctoral thesis is based on the publications listed above which have been reprinted with the kind permission of the publishers in the publications chapter on page 77 and which are referred to in the dissertation by their Roman numerals.

The personal contribution of the author of this thesis to the publications referred above is as follows:

- Paper I. Author of the paper. Leading the experiment, development and implementation of a pre-ingest tool described in the paper.
- Paper II. Author of the paper. Designing a new approach for defining designated user communities for archives. Author of the TRIANGULAR profiling method.
- Paper III. Author of the paper. Participating in the design, planning and co-leading the development of the archival access portal.

Presenting the work at the MTSR 2013: 7th Metadata and Semantics Research Conference.

- Paper IV. Main author of the paper. Participating in the design, planning and co-leading the development of related tools.

Disclaimer: The expressed content of this dissertation represents the view of the author, but not necessarily the view of any sponsor or other stakeholder.

Part of this work (E-ARK project-related activities) has been funded by the European Commission under its ICT Policy Support Programme (PSP) within the Competitiveness and Innovation Framework Programme (CIP).

1. INTRODUCTION

"*They have computers, and they may have other weapons of mass destruction.*"
(Janet Reno)

1.1. Background and previous research

The study conducted by the International Data Corporation (IDC) in 2014 revealed that there was approximately 4.4 ZB (zetabytes) of born-digital or digitized information created or replicated in 2013 (EMC Digital Universe, IDC 2014). This is a huge challenge to preservation organisations, as the growth of information is so fast, and existing volumes are so large and varied. Of course, not all of this information has permanent or long-term value, but some of it definitely has, and therefore should be kept for future generations. The fact that digital preservation is important but difficult to implement is well known among memory institutions. In 2009 the Planets project conducted a survey about the awareness of digital preservation. The project researched 158 archives, libraries, and other public sector agencies (including governmental agencies). The outcome was that 93% of respondents were aware of the challenges regarding digital preservation, but only 25% had a digital preservation solution (Tessella, British Library 2009). Such a drastic difference is presumably caused by the fact that digital information has several complex facets which need to be considered and which make the digital information vulnerable and its preservation resource consuming. Despite the fact that digital information is very fragile and complex, the memory institutions should still think ahead and guarantee the accessibility, usability, and comprehensibility of important digital information over time for the society.

1.1.1. Accessibility/availability

Firstly, the technologies used in the digital world are usually quite complex and require special conditions and handling skills. When comparing the situation with the paper world, then it can more easily occur that mishandling or unsuitable work conditions of the technology can do something irreversibly destructive with the information carrier and/or with the hardware. For example, if a compact laser disk (CD) gets an 8-micron (0,008 mm) scratch on the top side of the disk, then the disk or part of it can be corrupted as the information written to the CD is only covered by a 6–7-micron-deep layer of lacquer (Mueller 2004). Another example regarding the fragility of digital technology can be demonstrated by the hard drive, as many of the sophisticated electronic components can be damaged by charges as low as 10 volts (IT Business Edge 2003). Giving a 10V charge to a hard drive could harm the information on it by changing the polarity of bits or corrupting the drive completely. It is fairly

simple for humans to produce 10V, as ordinary static charge (e.g. produced by touching your hair) is at least 1500V (Andrews 2009). When we look at the analogous examples from the paper world, then there is quite a high probability that conservators can repair similar (0.008 mm) micro-scratches on paper sheets. With regard to electrical charge, a 10V charge applied to a sheet of paper will not affect the accessibility of the text on that sheet in the slightest.

As digital information must always have some physical carrier (e.g. magnetic disk), all physical dangers (e.g. fire, fluid, etc.) that are valid in the analogue world are issues in the digital world as well.

Furthermore, just as ink becomes difficult to read and paper gets brittle as time passes, similar changes also occur in the digital world. For example, the lifetime for CDs is approximately 5–59 years (Rothenberg 1999, pp 4). As CDs (in the form as we know them currently) were invented in the early 80s, it is obvious that the previously mentioned 59-year limit does not represent the real experience, but rather the estimated calculations and tests performed in laboratories. It also indicates that paper documents can be preserved even longer than CDs under the right physical conditions.

Although some information carriers are resistant to physical impact factors, they still become obsolete after some time. The reason is that the renewal cycle for technology is quite short (3–5 years) (Thomaz 2006). For example, many people have 3.5" magnetic disks (widely known as floppy disks) which were popular 10–15 years ago, but they do not have any drives / readers for those disks. As the market already provides disks which can store more information, it is not feasible for companies to produce, update, or support anything for old technologies. The obsolescence of drives / readers and information carriers is a difficult problem which does not have any other good solution except for a constant technology watch and refreshment of hardware, including information carriers (e.g. copying information from CD to DVD). It is recommended as reasonable to select the right information carrier (marked as long-term suitable) when creating the information, but we must admit that technology evolves very fast and new information carriers with better characteristics (more capacity, more persistent, etc.) are entering the market continuously and we then need to copy the content again and again to the newer media.

Perhaps the most radical solution is building computer museums as a solution for this problem. This means that along with the information carriers the original hardware and software will also be preserved. This solution is very expensive, as the collection of the technology grows very fast in time. The growth is even particularly expensive in case the equipment is preserved in duplicates to try to minimize the risk of equipment failures. Experts usually reach the same conclusion: building computer museums for digital preservation is not sustainable (Borghoff 2006).

Still, guaranteeing the accessibility to archived digital information is quite reasonably covered at memory institutions by following decentralized, geographically disparate preservation models like LOCKS (Lots of Copies Keep Stuff Safe) or similar. For example, the National Archives of Estonia (NAE) has

decided to reduce previously described risks by having the same information saved in several different geographical locations and by using different archival (long-term suitable) information carriers (avoiding a reoccurrence of the same information carriage failure twice) (Rahvusarhiiv 2008a). Of course, choosing the right physical conditions for the hardware and storage media and employing skilled people is a priority as well for the NAE.

1.1.2. Usability/representability

Usability can be defined as the ability for the user to do something sensible with the bits (Giaretta 2007). It actually means that we will need to deal with another big issue in the digital world, which is that the physical form of a bit of digital information does not define its precise logical form. For example, if we look at a CD with a very good microscope, then we will see the pattern which the laser has produced during the saving process (see Figure 1). The algorithm used for saving bits defines that all beginnings and endings of the "hole" are interpreted as ones (it also means that the "hole" is not exactly 1) and untouched space as zeros. This means that the holes do not reflect the value of a bit or a state of one bit.

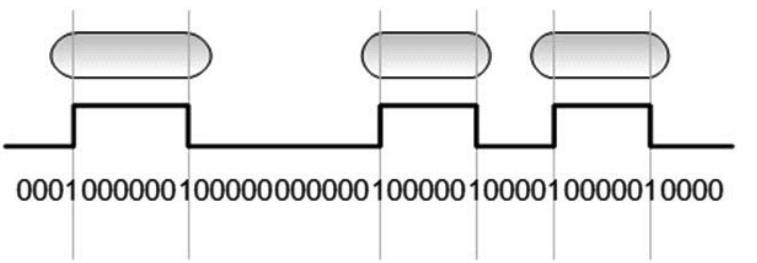


Figure 1. Bits saved to a compact disk

Thus, to save and read the information, we need to know the exact algorithm. In this case we need to know that when reading the disk we add 1 to the bit stream when encountering the edge of the hole, all other bits are marked as zeros. This therefore means that we need to have drivers for interacting with input devices, and when we have read bits to the memory, then we will need to have operation software and application programs for implementing bit streams.

However, bit streams can be implemented in many ways, as it all depends on what logic and algorithms we use. As we can interpret them in different ways and get different representations, it would be very difficult to decide what representation is the right one as shown in Figure 2 (Rothenberg 1999, pp 7).

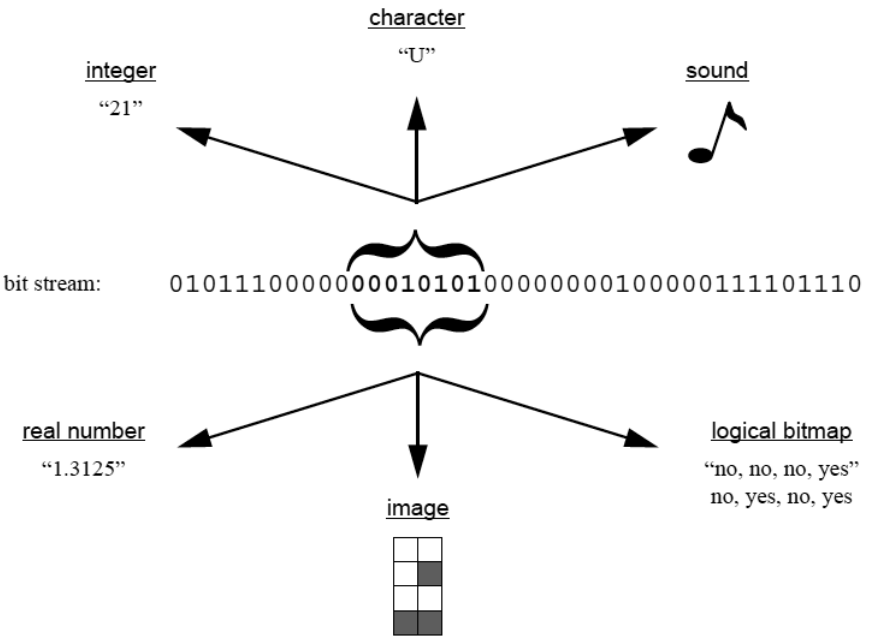


Figure 2. Bit stream with different representations (Rothenberg 1999, pp 7)

When we have information about the organising logic of bits, then we know what code table to use, etc. and we can create the original representation.

This also already involves the file formats – coding is always related to data types and application software. There can be many data types (text, maps, videos, audio, graphs, formulae, etc.). Each type can have many different codings. It is normal that some new formats will arise and some old formats become obsolete over time. This also means that when a user starts to use some new operating system, it may be difficult to find the right software to open some old files.

There are two main solutions to that problem: emulation and format migration. The InterPARES project defines emulation as *the reproduction of the behaviour and results of obsolete software or systems through the development of new hardware and/or software to allow execution of the old software or systems on future computers* (Interpares 2016). The glossary of the National Digital Stewardship Alliance defines emulation as *a means of overcoming technological obsolescence of hardware and software by developing techniques for imitating obsolete systems on future generations of computers* and (format) migration as *a means of overcoming technical obsolescence by preserving digital content in a succession of current formats or in the original format that is transformed into the current format for presentation. The purpose of format migration is to preserve the digital objects and to retain the ability for clients to retrieve, display, and otherwise use them in the face of constantly changing technology* (NSDA 2015). Both digital preservation strategies are used among

memory institutions and each approach has its own advantages and disadvantages, as seen in Table 1.

Table 1. Emulation vs migration (Stuchell 2013)

Emulation	**Migration**
• Can retain the 'look and feel' of the original digital object.	• Can retain the 'look and feel' of the original digital object, depending on migration strategy as well as the format being migrated to.
• The focus is on recreating the experience, not just accessing the content.	• May lose original formatting, causing the object not to look quite the same as it did.
• Preserve and use the original digital object.	• The focus is on making the content available.
• The emulator will also have to be preserved + will need to be updated periodically.	• May or may not save the original digital object for backup/future migration purposes. The file made in migration is a new copy.
• Without the original hardware/software, you can't be sure you're retaining the exact look and feel.	
• Does not always result in a perfect presentation of the original digital object.	

As the emulation does not change the original object (the bits remain the same), the hardware and software around it should be developed. Building such an emulator is very expensive work and validating the outcome can be complex as well, as this process still requires some human intervention. Still, many complex objects that archival formats are unavailable are good candidates for emulation. The reason is that the migration strategy cannot help to solve the situation in a sufficient way, as the next suitable long-term file format may just not exist (e.g. computer games). Even so, each time the decision point arrives, it is reasonable to check whether the next (long-term) format exists which is widely used, with an open source code, widely supported by software, and standardized. Constant technology watch and updating the list of suitable archival formats is one important part of the solution.

Both emulation and migration help preserve digital objects, but at the same time leave open several questions. One of them is how to guarantee understandability/comprehensibility next to usability. There is a need to be able to preserve the understandability and usability of the information encoded in digital objects (Giaretta 2011).

1.1.3. Understandability/comprehensibility/intelligibility

Understanding digital information first requires a representation of the archived information, so that a user could then interpret and understand it. However, it is a not-so-well-known fact that the digital information does not have any fixed representation before involving some software. For example, if we create a document in WordPad version 6.1 software, as seen in Figure 3, and save it as file "test.rtf", then it is obviously understandable for human users.[1] The user can probably see one table, then one image with the text and frame around them, and a couple more text rows below the frame.

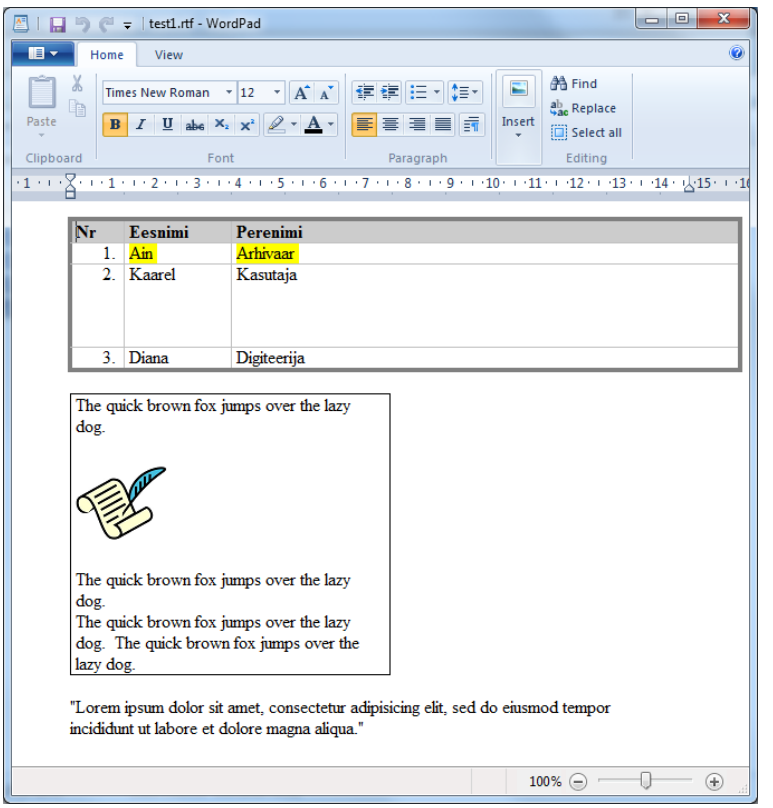


Figure 3. The representation in WordPad version 6.1

If we open the same "test.rtf" file in Hex Editor Neo software, then we will see the binary representation which is also correct but not suitable for human users, as humans are not used to interpreting binary codes (see Figure 4).

[1] Let us consider the language of the document irrelevant in this example.

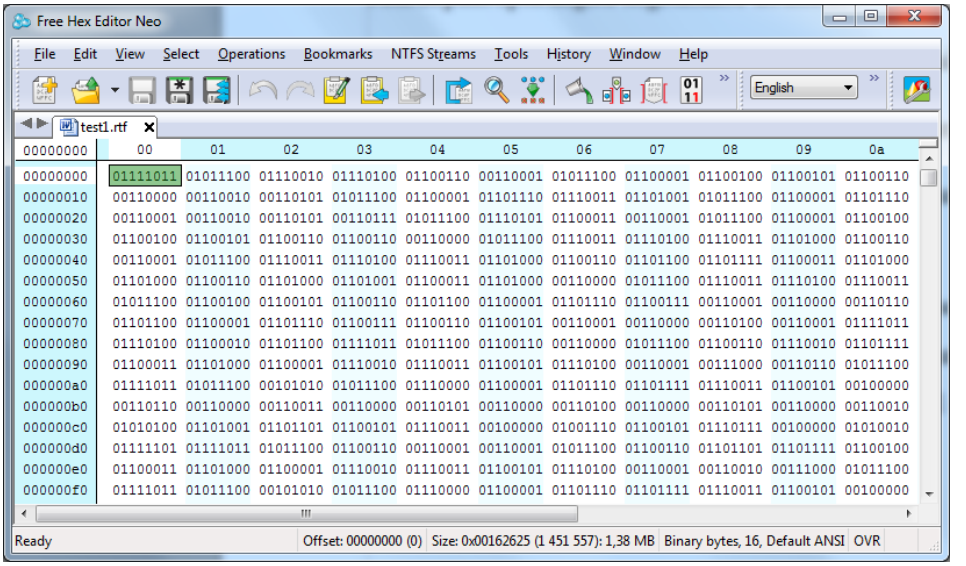


Figure 4. The representation in Hex Editor Neo version 5.10

When we open that file in Notepad++ version 6.1.1, then we can see the structure of the RTF coding. Again, this is the correct interpretation of this file, but not understandable for the ordinary user, as it shows the technical view of the file format structure (see Figure 5).

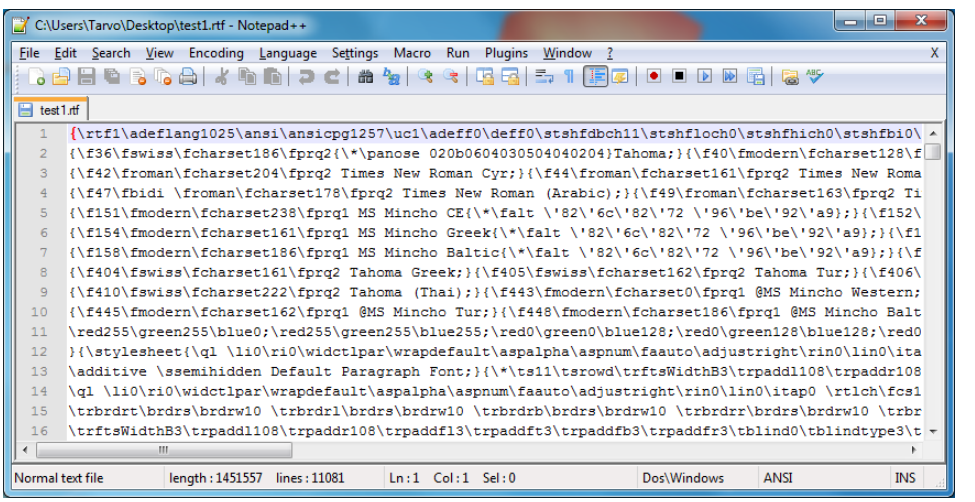


Figure 5. The representation in Notepad++ version 6.1.1

When we open that file in Microsoft Word 2010, then there are also no errors displayed, but we can notice some differences (lines between rows are missing, the table has two borders) compared to the original WordPad version 6.1 representation (see Figure 6).

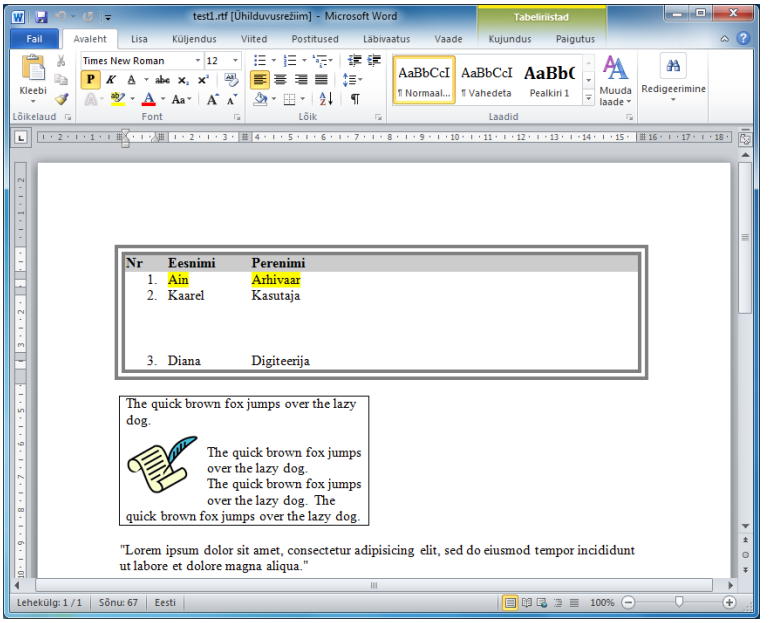


Figure 6. The representation in Microsoft Word 2010

If we open this file in OpenOffice Writer (see Figure 7) or LibreOffice Writer (see Figure 8), then we will notice even more changes, although the original bits are the same and no errors are displayed by the software.

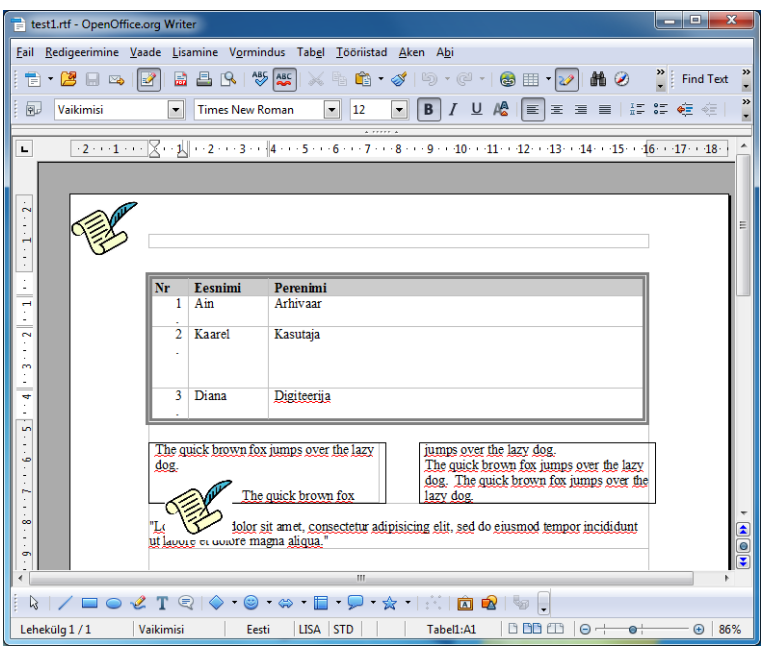


Figure 7. The representation in OpenOffice Writer 3.3.0

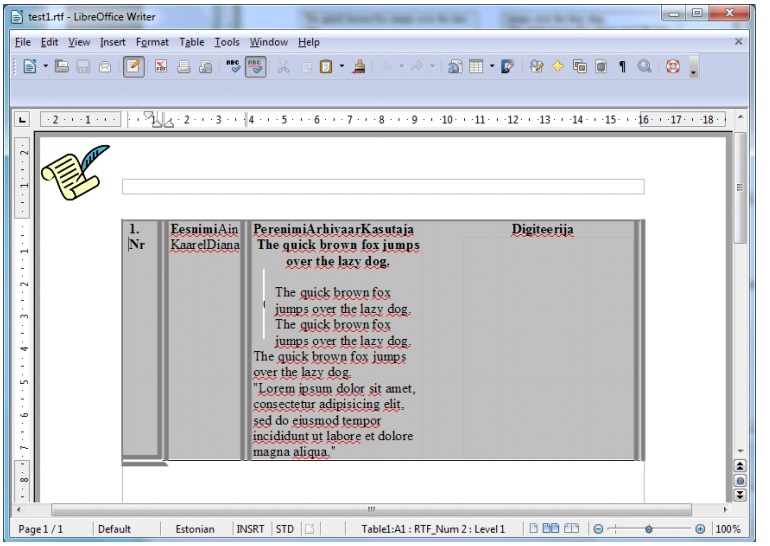


Figure 8. The representation in LibreOffice Writer 3.4.4

Thus, all representations are technologically correct and no errors will be displayed to the user when they are opening this file. It is quite obvious to the human eye without even seeing the first representation that something is wrong with the representation in the last two examples (Figure 7 and Figure 8), but there are many cases when this is not so clear. For example, let us look the table below (Table 2).

Table 2. Background colours can represent significant properties

A

1400	200	300	1100
400	1000	500	1100
800	900	400	100
800	600	700	200
1800	1900	2600	4200

B

1400	200	300	1100
400	1000	500	1100
800	900	400	100
800	600	700	200
1800	1900	2600	4200

Let us agree that the first view (A) is the original representation. The cells are coloured because they represent values which will be summed at the end of each column (for example, all yellow cells give the answer 1800=1400+300+100). If we open that table in some other software which cannot represent background colours or the colours go missing during format migration, then we get the table on the right (B), which does not make much sense to the human user.

To conclude, the examples have shown how we can say that digital information totally depends on technology – we need technological tools for mediating creation, preservation, and access, as shown in Figure 9.

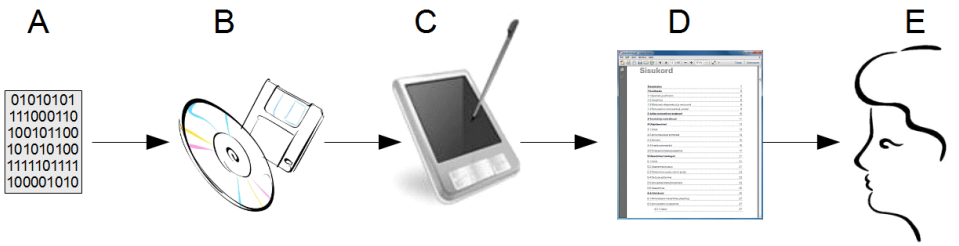


Figure 9. Digital information depends on several factors

The digital information is encoded in binary form (A), but it cannot exist without some physical form (B). The information carrier requires some hardware and drivers to make a connection between the bits on the information carrier and the system memory (C). The digital information needs some operating system and other software (D) to give a form to the digital information that is understandable to humans. The user can then access that information (step E). Thus, the perception step depends on how authentically the information is presented to the user, but also what knowledge base the user has. This is the most crucial step in this access workflow.

Let us look at another example to support the importance of this statement. If in some information system we see the date 10/12/09 and there is no information about the date format, then in the Estonian cultural space it usually represents 10 December 2009. In USA it may be seen as 12 October 2009 and in China 9 December 2010. This therefore reflects that besides the technically authentic representation, we should strongly consider the context and what knowledge a user may have, as users of different backgrounds and knowledge bases may sense the information very differently.

The Open Archival Information System (OAIS) reference model (see also page 28), which is widely adopted all over the world by all kinds of archival organizations, describes a knowledge base as *a set of information, incorporated by a person or system, that allows that person or system to understand received information* (CCSDS 2012, pp 22). From the perspective of memory institutions, it means that they will need to start identifying and getting to know their user groups, in order to start helping users to understand the preserved information. It also means that when memory institutions are developing new access solutions, they should definitely incorporate these components and descriptions which are familiar to users of their solutions. However, it is important to realise that the user communities are not going to stay the same as time passes; on the contrary, the designated user communities are live organisms which will change over time. According to OAIS, a Designated Community is *an identified group of potential Consumers who should be able to understand a particular set of information; it may be composed of multiple user communities; it is defined by the archive and this definition may change over time* (CCSDS 2012, pp 21).

If we assume that in the access phase, the memory institutions can provide information that conforms to the knowledge which users possess, then it is

possible to lead different users to the same or similar perceptions, although they had different knowledge bases when they started the access process. This means that it is very important which search and navigation functionalities of an access system are provided to users.

For example, let us assume that we have three users: a 1st year Estonian student at University of Tartu, a foreign student from a student exchange programme, and a high school student.

They all reach the archival access portal for some reason.

- The 1st year student wants to find information about the Swedish king who founded the University of Tartu in 1632. S/he does not remember the name of the king, and therefore inserts his current knowledge in the search form – s/he puts University of Tartu in the free text search box and selects the year 1632 as the time limit. S/he gets few results, but they all contain the king's name: Gustav II Adolph.
- The foreign student does not have such detailed information about the university, but wants to learn about the history of the university and therefore s/he first changes the portal language to English (as s/he hails from the English cultural space) and selects the organisation, Tartu University, from the organisation list which contains all organisations that have something archived. From the description s/he can easily find that the University of Tartu was founded by King Gustav II Adolph in 1632.
- The high school student wants information about King Gustav II Adolph for their essay and s/he therefore opens the people search section on the access portal's front page and inserts the king's name. S/he gets few results, and sees from there that the same king was also the founder of the University of Tartu and that the university was founded in 1632.

All users had different knowledge bases, but despite that they all reached the same results in the end. This means that by combining multifaceted search options we can influence users' knowledge bases to sense the information in an objective way.

As seen from the previous simple examples, knowledge definition is actually two-fold. The Open Archival Information Model (OAIS) also reflects the same implication in stating that knowledge can be incorporated by a person or system as described earlier – it means that a knowledge base can exist both in a human and/or in a system/machine. When we look next to the definition of information, then OAIS defines it as *any type of knowledge that can be exchanged* (CCSDS 2012, pp 22). When we compare these two definitions we realise that we can end up in a tautological cycle when trying to make a very clear distinction between knowledge and information – by the OAIS definitions, knowledge can be information and information can be knowledge.

Perhaps a clearer definition about the system-incorporated knowledge has been composed by Reagan W. Moore. He introduces the term "digital entity", and explains that these are *images of reality, stored as data (bits) that must be interpreted for display by applying information (semantic tags that provide the*

meaning of the bits) and knowledge (structural relationships defined by a data model) (Moore 2001).

To avoid further misunderstanding and confusion, a fixed definition of "knowledge" must be provided. There are numerous classifications of knowledge focusing on different dimensions of various fields (philosophy, business, religion, etc.) available, but taking into account the previous definitions presented here, we can conclude that in the sense of this dissertation we define knowledge in two ways (EKMPT 2010):

- tacit knowledge (stored in people's heads);
- explicit knowledge (printed matter and electronically stored information).

According to Nonaka, not all the tacit knowledge incorporated by a person can be transformed to explicit knowledge (Nonaka 1994). A typical example of such tacit knowledge is the skill of riding a bicycle.

It is different with the electronically stored information. The digital information can be organized and put into the correct stored context by adding relations and links. The links can be formal descriptions of the dependencies of digital information.

According to Moore, the digital knowledge could be represented by relationships defined by a data model (Moore 2001).

The knowledge residing in the archives' repository should be "independently understandable". According to OAIS the term "independently understandable" is a *characteristic of information that is sufficiently complete to allow it to be interpreted, understood, and used by the Designated Community without having to resort to special resources not widely available* (CCSDS 2012, pp 22)

The comprehensibility of digital information depends on both – it is clear that to overcome the problems with accessibility, usability, and understandability the archives must take into account users' knowledge bases and structural relationships between information residing in systems. At the same time it is not very clear what theoretical and practical solutions can be implemented for that cause. The continuous understanding and use of digital objects is clearly identified as a basic requirement to the preservation process; however, the researcher and practitioner communities still have to define what this requirement means in the implementations of preservation systems (Dobreva 2009).

1.1.4. State-of-the-art

To approach the state-of-the-art in this topic a review of the current state of both research and practice is required. There are several ways to do that:

- One possible approach is to identify all major memory institutions and analyse their current practical solutions. Most significant research institutions in that field should be also identified and their work analysed to get the latest status of related research. To analyse all these organisations as

individual research objects is extremely resource consuming. For example, only the Archives Portal Europe alone lists already more than 200 archives from 33 countries (APE 2016). Therefore, this approach cannot be considered the best option for the current thesis.

- Another approach is to look at the situation from different perspectives and handle the memory and research institutions as groups via common elements between them. International digital preservation projects can be considered especially suitable for that kind of research, because:
  - they unite the target group (linking memory and research institutions);
  - the number of significant digital preservation projects is remarkably smaller than the number of individual organisations;
  - projects can reflect not only the past and the current status, but also help to predict future trends;
  - project analysis can point out different (both theoretical and practical) aspects of the current topic more effectively;
  - projects can be initiators for centres of excellence and foundations (e.g. Open Preservation Foundation) that sustain the research results.

To gain a better understanding of each relevant digital preservation project, separate overview tables (project overview cards) will be created. Each card is limited to about 2-3 pages in order to keep the overview focused. Each card has the following information:

- Name – clearly states the name of the project and acronym if available.
- Web page – presents the address to the projects website.
- Funded by – gives information about the main source of financing.
- Duration – gives information about the project timeline, and indicates whether the project has ended or is still ongoing.
- Partners – presents the list of all partners in the project and by such reflects the scale and the geographical coverage of the project.
- Description – summarises the objective(s) and scope of the project.
- Progress – describes the outcome of the project by listing all public deliverables.
- Relevance – analyses whether the project is relevant to this research or not.
- Note – provides space for any additional information about the project which is important to mention.

Disclaimer: The selection of projects does not claim to be complete. The selection represents the view of the author and is influenced mainly by the following factors:

- Coverage (investigating only international projects with more than 5 partners, focusing on European projects in order to keep the scope manageable).
- Subject relation (investigating only projects which seem to have at least a partial relation with the preservation of digital knowledge).

- Time (investigating only projects which are still ongoing or which have ended not a long time ago to avoid the obsolescence of information which is very common in information science and technologically advanced economics).
- Dissemination (investigating only projects which have produced at least some public deliverables).
- Continuity (creating project cards about new projects even after the start of the research if they are being discovered, although their results may be not fully reflected in the research)

The tables are available in "Appendix A. The full version of project cards"," but the information about relevance has also been extracted to the following Table 3 for better reading purposes.

Table 3. The relevance of digital preservation projects

No	Project Name	Relevance
1.	Collaboration to Clarify the Costs of Curation (4C)	The project was focused on estimating and comparing the cost of digital curation (as well digital preservation). Although the cost component is gathering more and more importance in the digital preservation area, the 4C project will not be further investigated as the curation cost modelling is out of the scope of the research planned for this dissertation.
2.	Alliance Permanent Access to the Records of Science in Europe (APARSEN)	The project was focused on identifying and merging research areas of digital preservation into a common vision. APARSEN addresses solutions to problems in digital preservation through consultancy, training, tools and services. As the dissertation deals with more detailed and focused research questions (more specific to digital preservation of knowledge) then the project will not be further investigated.
3.	ARchive COmunities MEMories (ARCOMEM)	The project was focused on social web, automated information creation and appraisal. Some reusable software components to support the collection, analysing and access of content from Web and Social Web have been delivered. Although the project took the leverage of wisdom of crowds, it still has a different focus compared to this research. Therefore, as the dissertation deals with research questions more specific to digital preservation of knowledge, then the project will not be further investigated.

No	Project Name	Relevance
4.	BlogForever	The project was focused on harvesting, preserving, managing and reusing blog content. More specifically, the system provided by BlogForever retrieves, parses hypertext and associated content (e.g. images) from blogs and is then carving semantic entities (e.g. author names) out of the parsed content. Although the tools and methods provide good means to harvest semantic descriptions about the content, they have still too strong focus on blogs. As this dissertation seeks solutions for more universal content, then the BlogForever project will not be further investigated.
5.	Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval (CASPAR)	As CASPAR aimed to investigate not just the bits of digital objects, but also the information and knowledge preservation then it belongs to the interest group for this research. It is also very important that the CASPAR project brought together a consortium covering extensive scientific, cultural and creative expertise in the field of information preservation. The results of the project will also give a very good overview of issues related to the representation information. Therefore it is most reasonable to investigate that project more closely when finding answers to the research questions of this dissertation. Especially valuable is the concept of profiling a Designated Community and formalizing intelligibility by using modules and dependencies. This concept has been explained in deliverables D2101 (Giaretta 2008), D2101B (Tzitzika et al. 2008, 1a) and D2102 (Tzitzika et al. 2008, 1b) of the CASPAR project. Therefore, these deliverables will be the most relevant of CASPAR deliverables for this research. As profiling designated communities is an important part in digital knowledge preservation area, the topic will be further investigated in the light of article II.
6.	European Archival Records and Knowledge Preservation (E-ARK)	The E-ARK project addresses several important issues of digital preservation of knowledge. Most important are pre-ingest and ingest activities (including the specification of suitable format for a submission information package). The author of this dissertation is also the work package lead of WP3 “Transfer of records to archives” in E-ARK project which eases the research process as a significant amount of specific knowledge is available to use. The project has more than 15 partners from different organisations (national archives, research organisations, vendors, universities etc.). One of the partners is the DLM Forum which brings to the E-ARK project skills and knowledge from 22 national archives.

		Most relevant deliverables from the perspective of this research are: • D3.1 Report on available best practices (Kärberg et al. 2014). This deliverable introduces best practices in digital archiving and pays special attention to records export, archival ingest workflows, submission information package formats and ingest of digital objects and their metadata. This report is significant as it helps to understand how memory institutions and other stakeholders deal with digital objects, what is common and what is not. It also reflects that there is no common format for submission information packages (SIPs) or even in some cases clear understanding of what a SIP is, although the term is well described in the OAIS model. • D3.2 E-ARK SIP draft Specification (Kärberg et al. 2015). This deliverable describes a draft SIP specification for the E-ARK project and provides initial input for the technical implementations of E-ARK ingest tools. As the format is based on available best practices, it is a good candidate for a common central format for SIPs all over the world. • D3.3 E-ARK SIP pilot specification (revision of D3.2) (Kärberg et al. 2016). This deliverable extends the deliverable D3.2 by providing a revised version of its content, adding more details relevant for tool development and implementation, and describing specific profiles for the transfer of relational databases, electronic records management systems (ERMS) and simple file system based records (SFSB). As the author of this thesis was also actively involved in the writing process of the before mentioned documents, then there was good alignment of the relevant outcomes of the E-ARK project and this thesis.
7.	Enabling kNowledge Sustainability Usability and Recovery for Economic value (ENSURE)	The project explored issues specific to digital preservation in the context of the industry and services sector. As the project addressed mainly the long-term usability of commercially relevant data produced or controlled by organisations, then it does not belong to the scope of this research.
8.	ForgetIT. Concise Preservation by combining Managed Forgetting and Contextualized Remembering.	The project was focused on digital preservation issues of organisational and personal knowledge. As the project started rather late (2013) compared to the time schedule of this research, and as this project is mainly focused on organ-isational (excluding memory institutions such as national libraries and archives) and personal knowledge preservation, then it will be not further investigated in the scope of this research.

9.	Keeping Emulation Environments Portable (KEEP)	The project was investigating various aspect of emulation, in theory and in practice, and has prompted many organisations to actually use emulation for the first time. It also provided tools etc. that have led to Emulation as a Service, used now by several memory institutions. Although enabling access to and use of digital objects stored on outdated computer media is important, the main focus of this project is still out of scope of the research carried out in this dissertation. Therefore, the project will not be further investigated.
10.	Living Web Archives (LIWA)	The project was addressing Web Archiving. The tools developed and demonstrated by this project allow capturing web content from a wide variety of sources. Although LiWA delivered two exemplary applications for demonstrating the benefits of advanced Web archiving to interested stakeholders, they still fell out of this selection as they are focusing on audio-visual streams and social web content respectively, but this dissertation seeks solutions for more universal content.
11.	Preservation and Long-term Access through Networked Services (PLANETS)	The PLANETS project addressed mainly preservation planning, characterisation of digital objects and automation with scalability. It was a huge and influential project – the project delivered a comprehensive framework and a number of practical tools for digital preservation and access. As the project was more focused on the actions that happen inside an archive (effective preservation planning, preservation characterisation techniques, preservation actions, etc.) rather than ingest or pre-ingest then the PLANETS project is not very strongly related to this thesis. Although some ideas introduced in PLANETS (especially in the preservation panning tool PLATO) will influence the definition of archival user communities in the article II.
12.	PREserving Linked DAta (PRELIDA)	As the project aimed to build bridges across the digital preservation and linked data communities (making the linked data community aware of existing outcomes of the digital preservation community) then it is also an important project for this dissertation. Although the dissertation is not focused on linked open data, it is still part of the knowledge preservation and will be further investigated in the light of article IV.
13.	PrestoPRIME	The project was addressing long-term preservation of and access to digital audio-visual content by integrating media archives with European on-line digital libraries. Although the project developed a number of technical solutions for managing the preservation content, it is not strictly in the scope of this research, as this research targets a more universal approach to the content – the dissertation deals with research questions which are not audio-visual content specific. The project will not be further investigated.

14.	PReservation Organizations using Tools in AGent Environments (PROTAGE)	The project built and validated software agents for long-term digital preservation and access that can be integrated into existing and new preservation systems. As the project looked for ways to coordinate the pre-ingest phase more effectively then the results of the PROTAGE project will be respectfully observed. A more detailed look to the documentation (NAE 2010) will be taken in the preparation phase of the article I of this research.
15.	Scalable Preservation Environments (SCAPE)	The project primarily addressed the scalability issue of digital preservation. SCAPE delivered scalability improvements on some existing tools (i.e. Plato), new scalable tools (i.e. Hawarp) and tools for quality control (i.e. Matchbox). As the research questions of this dissertation are more specific to digital preservation of knowledge then the SCAPE project will not be further investigated.
16.	Sustaining Heritage Access through Multivalent Archiving (SHAMAN)	The project developed tools for analysing, ingesting, managing, accessing and reusing information objects and data across memory institutions. It also explored the ways of using GRID computing in digital preservation. As there is not enough public information available about the outcomes of this project then it cannot be further investigated.
17.	Timeless Business Processes and Services (TIMBUS)	The project explored issues specific to digital preservation of enterprise business processes. As the project addressed mainly the long-term usability of commercially relevant data processes (aligning with the enterprise risk management and business continuity management) then it does not belong to the scope of this research as this thesis is focused on the knowledge preservation issues related to memory institutions.
18.	Advanced Workflow Preservation Technologies for Enhanced Science (Wf4Ever)	The project aimed at providing the methods and tools required to ensure the long-term preservation of scientific workflows. As the dissertation deals with more knowledge-oriented research questions (more specific to digital preservation of knowledge and not science-centric) then the project will not be further investigated.

As seen from the overview cards, many digital preservation projects have researched different aspects of knowledge preservation. Some projects are focused on

- some very specific issue/topic (4C, KEEP),
- issues in some specific sector (ENSURE, ForgetIT, TIMBUS, Wf4Ever)
- on web preservation issues (LIWA, BlogForever)
- interoperability and information semantics issues (APARSEN, CASPAR, E-ARK, PRELIDA)
- pre-ingest activities (E-ARK, PROTAGE)

- ingest issues (ARCOMEM, BlogForever, LIWA, E-ARK)
- automation and preservation issues inside an archive (PLANETS, PrestoPRIME, SCAPE)
- full OAIS (SHAMAN)

Based on the state-of-the-art analysis, the most relevant for this thesis is the European Archival Records and Knowledge Preservation (E-ARK) project, as it has the strongest relations with the scope of this research. Although this dissertation is not entirely based on the E-ARK project, it still has an important role to fill in this research. More precisely, the E-ARK project will help to:

- identify, collect and investigate European best practices for archiving via deliverable D3.1 E-ARK Report on Available Best Practices[2];
- develop a common (international) submission information package specification (SIP) for archival knowledge ingest (deliverables D3.2 E-ARK SIP Draft Specification and D3.3 E-ARK SIP Pilot Specification).

Best practices introduced in projects like CASPAR, PLANETS, PRELIDA, PROTAGE will be also taken into account in the way described in the Relevance row on the respective project cards.

1.1.5. Open Archival Information System (OAIS)

As the current thesis is very tightly related to the OAIS model, the model and the need for it should be explained to the readers.

What is the OAIS model?

The OAIS model is an ISO standard (14721:2012) and a recommended practice (650.0-M-2) from the Consultative Committee for Space Data Systems (CCSDS) which defines the reference model for an open archival information system. According to the magenta book of the Reference Model for an Open Archival Information System, it is a technical recommended practice for use in developing a broader consensus on what is required for an archive to provide permanent, or indefinite long term, preservation of digital information and it establishes a common framework of terms and concepts which make up an Open Archival Information System (CCSDS 2012, p 5).

Why is OAIS important for digital preservation?

The OAIS standard (CCSDS 2012, p 11)

- provides a framework for the understanding and increased awareness of archival concepts needed for long term digital information preservation and access;
- provides the concepts needed by non-archival organizations to be effective participants in the preservation process;

[2] This report will give even broader overview about the state-of-the-art in digital preservation.

- provides a framework, including terminology and concepts, for describing and comparing architectures and operations of existing and future archives;
- provides a framework for describing and comparing different long term preservation strategies and techniques;
- provides a basis for comparing the data models of digital information preserved by archives, and for discussing how data models and the underlying information may change over time;
- provides a framework that may be expanded by other efforts to cover long term preservation of information that is not in digital form (e.g. physical media and physical samples);
- expands consensus on the elements and processes for long term digital information preservation and access, and promotes a larger market which vendors can support; and
- guides the identification and production of OAIS-related standards.

In other words, the OAIS model provides a conceptual model which specifies how digital assets should be preserved for a community of users from the moment digital material is ingested into the digital storage area, through subsequent preservation strategies to the creation of a dissemination package for the end user (Johnston 2013).

Why is OAIS important for this research?
The OAIS model values the identification and application of appropriate preservation strategies to maintain archived digital objects in a readable, usable and understandable state. The model points out the importance of target consumers (understanding their interests, needs) and their significant implications for the preservation of ingested digital objects. The most significant OAIS requirements for a digital preservation repository (from the perspective of this dissertation) are

- to determine the designated user community (DC);
- to ensure that the archived objects are independently understandable to the DC;
- monitor the DC;
- develop packaging principles and procedures.

The OAIS model is important because it provides also a common vocabulary, a high-level data model and describes the required responsibilities and recommended functional entities for digital preservation.

1.2. Objectives

As previously described, digital preservation of knowledge is a very broad and complex research area, and this dissertation will therefore focus only on one major topic – researching continuous comprehensibility and meaningful use of

digital objects by taking into account the designated users' knowledge and OAIS model.

Therefore, all research presented in this dissertation will be strongly influenced by OAIS model core entities, as they will be the backbone of the research.

The aim is not to provide one and only solution to the research problem, but rather introduce and analyse the different aspects of the problem, develop solutions and test the outcomes in practice.

1.3. Research questions

The research questions will be grouped by the functional entities of the OAIS model. The OAIS model has six core functional entities: Ingest, Archival Storage, Data Management, Administration, Preservation Planning and Access.

1. Ingest Functional Entity

According to OAIS, the Ingest Functional Entity provides the services and functions to accept Submission Information Packages (SIPs) from Producers (or from internal elements under Administration control) and prepare the contents for storage and management within the Archive (CCSDS 2012, pp 44) as seen in Figure 10 (CCSDS 2012, pp 48).

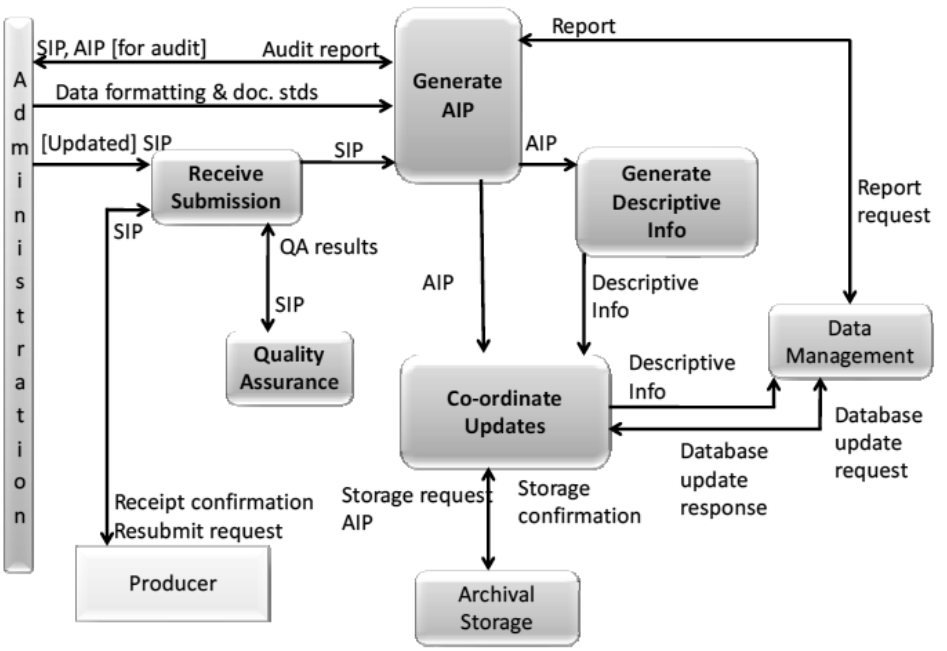


Figure 10. Ingest Functional Entity (CCSDS 2012, pp 48)

According to OAIS, the functions can be summarised as seen in Table 4 the cells with a cray background indicate relevant functions).

Table 4. Ingest Functional Entity

Name	**Description**	**Relevance to the research**
Receive Submission	Provides the appropriate storage capability or devices to receive an SIP from the Producer (or from Administration).	Is responsible for adequate Representation Information in the received submission.
Quality Assurance	Validates (QA results) the successful transfer of the SIP to the temporary storage area.	Allows double-checking as to whether all the necessary knowledge attributes in the SIP are in place.
Generate AIP	Transforms one or more SIPs into one or more AIPs that conform to the Archive's data formatting standards and documentation standards.	Does not include anything knowledge-specific and therefore is not observed in the research.
Generate Descriptive Information	Extracts Descriptive Information from the AIPs and collects Descriptive Information from other sources to provide Coordinate Updates, and ultimately Data Management.	This depends totally on the technical implementation and AIP format specifics. As it does not include anything knowledge-specific, it will be not observed in the research.
Coordinate Updates	Is responsible for transferring the AIPs to Archival Storage, and the Descriptive Information to Data Management.	Does not include anything knowledge-specific and therefore is not observed in the research. Updates and the enrichment of the archived information will be observed more closely in the Receive Data function in the Archival Storage Functional Entity.

It is clear that this entity not only contains functions for transferring the information to the archives, but also functions for enriching it. This allows the information received to be transformed to digital knowledge. Therefore, the main research question for complementing this entity is:

How can the ingest function be broadened to help prepare records for archiving already in the pre-ingest stage by adding detailed metadata about the structure, context, and relationships to transform the information to digitally codified knowledge?

2. Archival Storage Functional Entity

According to OAIS, the Archival Storage Functional Entity provides the services and functions for the storage, maintenance, and retrieval of AIPs (CCSDS 2012, pp 45) as seen in Figure 11 (CCSDS 2012, pp 51).

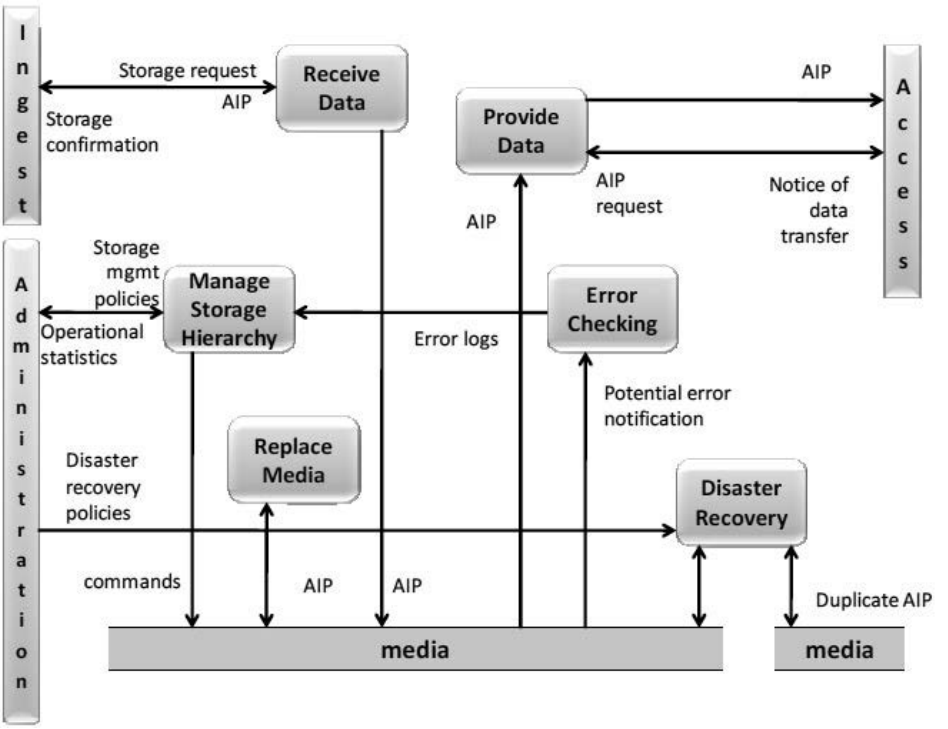


Figure 11. Archival Storage Functional Entity (CCSDS 2012, pp 51)

According to OAIS, the functions can be summarised as seen in Table 5 (the cells with a grey background indicate relevant functions).

Table 5. Archival Storage Functional Entity

Name	Description	Relevance to the research
Receive Data	Receives a storage request and an AIP from Ingest and moves the AIP to permanent storage within the Archive.	Allows enrichment of the archived information. Actual enrichment is coordinated by the Archival Information Update function in the Administration Functional Entity.
Manage Storage Hierarchy	Positions, via commands, the contents of the AIPs on the appropriate media based on storage management policies, operational statistics, or directions from Ingest via the storage request.	Does not include anything knowledge-specific and therefore is not observed in the research.
Replace Media	Provides the capability to reproduce the AIPs over time.	Does not include anything knowledge-specific and therefore is not observed in the research.
Error Checking	Provides statistically acceptable assurance that no components of the AIP are corrupted in Archival Storage or during any internal Archival Storage data transfer.	Does not include anything knowledge-specific and therefore is not observed in the research.
Disaster Recovery	Provides a mechanism for duplicating the digital contents of the Archive collection and, for example, storing the duplicate in a physically separate facility.	Does not include anything knowledge-specific and therefore is not observed in the research.
Provide Data	Provides copies of stored AIPs to Access.	Does not include anything knowledge-specific and therefore is not observed in the research.

This entity contains functions for information storage in the repository. It also supports functions for enriching the stored digital information semantically (for transforming the archived information to digital knowledge by improving the quality of archived information) by the Administration Functional Entity.

As the Archival Storage Functional Entity is focused on storing Archival Information Packages (AIPs) and it is not knowledge-specific, then this entity will be studied together with the Administration and Data Management functional entities to cover the stored packages' metadata enrichment. See the next item for the main research question to this entity.

3. Data Management Functional Entity

According to OAIS, the Data Management Functional Entity provides the services and functions for populating, maintaining, and accessing both Descriptive Information which identifies and documents Archive holdings and administrative data used to manage the Archive (CCSDS 2012, pp 45) as seen in Figure 12 (CCSDS 2012, pp 53).

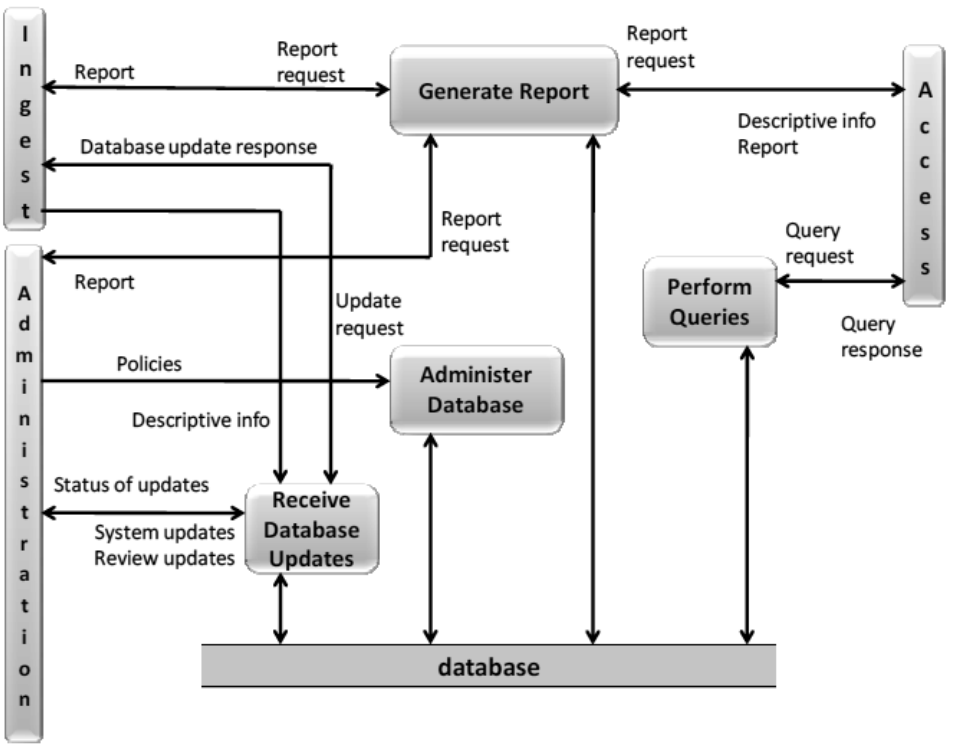


Figure 12. Data Management Functional Entity (CCSDS 2012, pp 53)

According to OAIS, the functions can be summarised as seen in Table 6 (the cells with a cray background indicate relevant functions).

Table 6. Data Management Functional Entity

Name	Description	Relevance to the research
Administer Database	Is responsible for maintaining the integrity of the data management database, which provides a storage mechanism that can be queried in some way for storing both Descriptive Information and system information.	Does not include anything knowledge-specific and therefore is not observed in the research.
Perform Queries	Receives a query request from Access and executes the query to generate a query response that is transmitted to the requester.	Does not include anything knowledge-specific and therefore is not observed in the research.
Generate Report	Receives a report request from Ingest, Access, or Administration and executes any queries or other processes necessary to generate the report that it supplies to the requester.	Does not include anything knowledge-specific and therefore is not observed in the research.
Receive Database Updates	Adds, modifies, or deletes information in the Data Management persistent storage. The main sources of updates are Ingest, which provides Descriptive Information for the new AIPs, and Administration, which provides system updates and review updates.	Allows for enriching the archived information.

This entity contains functions for updating storage (including functions for enriching the archived information). It can help to enrich the archived information in the storage to raise it to a new digital knowledge level in cooperation with the archival storage entity.

The Data Management Functional Entity does not involve anything knowledge-specific, but it provides means for updating archived knowledge and therefore it will be studied together with the Archival Storage Functional Entity in this thesis as explained previously.

The main research question for complementing this and the Archival Storage Functional Entity is the following:

How can users share their knowledge with archives?

4. Administration Functional Entity

According to OAIS, the Administration Functional Entity provides the services and functions for the overall operation of the Archive system (CCSDS 2012, pp 45) as seen in Figure 13 (CCSDS 2012, pp 54).

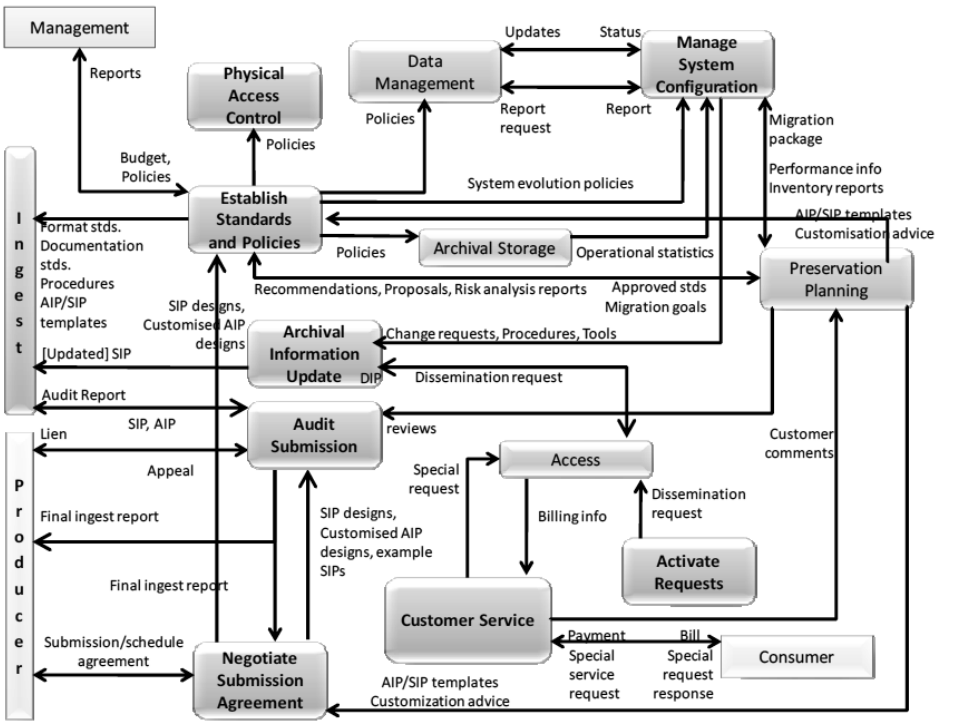


Figure 13. Administration Functional Entity (CCSDS 2012, pp 54)

According to OAIS, the functions can be summarised as seen in Table 7 (the cells with a grey background indicate relevant functions).

Table 7. Administration Functional Entity

Name	Description	Relevance to the research
Negotiate Submission Agreement	Solicits desirable archival information for the OAIS and negotiates Submission Agreements with Producers.	Allows for agreement on submission workflow and a data model which would be suitable for archiving knowledge.
Manage System Configuration	Provides system engineering for the Archive system to monitor the functionality of the entire Archive system continuously and systematically control changes to the configuration.	Does not include anything knowledge-specific and therefore is not observed in the research.
Archival Information Update	Provides a mechanism for updating the contents of the Archive.	Allows for updating the archived information (important for the enrichment process). The process is facilitated by the Ingest and Archival Storage functional entities.
Physical Access Control	Provides mechanisms to restrict or allow physical access (doors, locks, guards) to elements of the Archive, as determined by Archive policies.	Does not include anything knowledge-specific and therefore is not observed in the research.
Establish Standards and Policies	Is responsible for establishing and maintaining the Archive system standards and policies.	Allows for establishing policy for Designated Community management and standards for submission information packages.
Audit Submission	Will verify that submissions (SIP or AIP) meet the specifications of the Submission Agreement. In the case of the SIP and in the case of the AIP it verifies the understandability by the Designated Community.	Allows for validating submissions to be sure that everything needed for the Designated Community are represented in SIP.
Activate Requests	Maintains a record of event-driven requests and periodically compares it to the contents of the Archive to determine if all needed data is available.	Does not include anything knowledge-specific and therefore is not observed in the research.
Customer Service	Will create, maintain, and delete Consumer accounts.	Does not include anything knowledge-specific and therefore is not observed in the research.

As the Administration Functional Entity coordinates all the other entities in OAIS, then it is also considered in some detail together with other entities. For example, it supports specifying submission agreements, which is an important part of (pre-)ingest and therefore it will be studied together with the Ingest Functional Entity in this thesis.

5. Preservation Planning Functional Entity

According to OAIS, the Preservation Planning Functional Entity provides the services and functions for monitoring the environment of the OAIS, providing recommendations and preservation plans to ensure that the information stored in the OAIS remains accessible to, and understandable by, the designated community over the long term, even if the original computing environment becomes obsolete (CCSDS 2012, pp 45) as seen in Figure 14 (CCSDS 2012, pp 57).

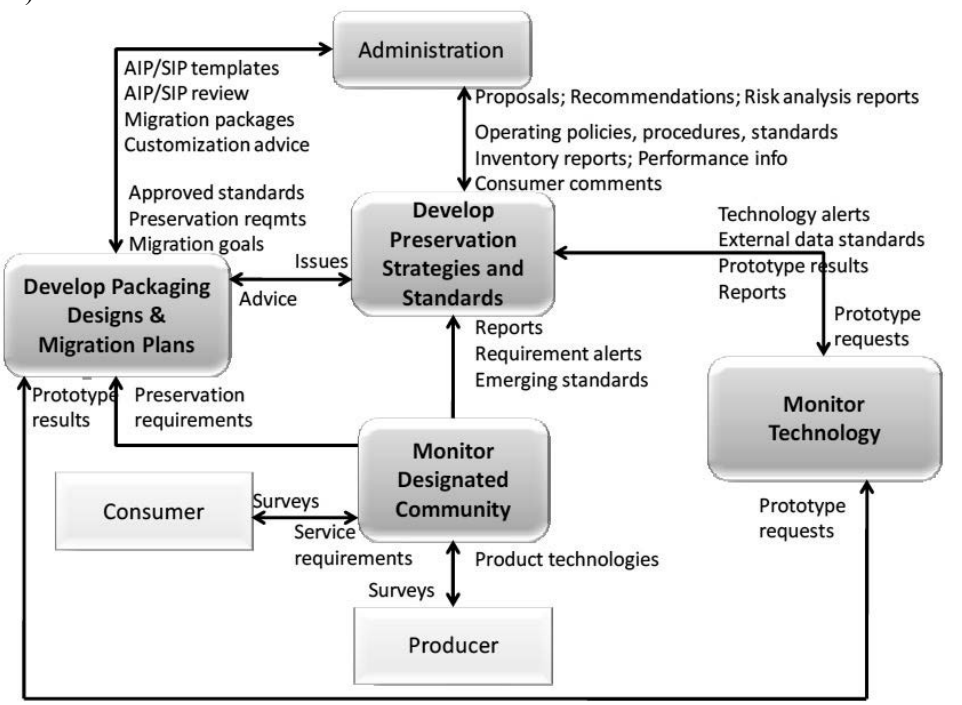


Figure 14. Preservation Planning Functional Entity (CCSDS 2012, pp 57)

According to OAIS, the functions can be summarised as seen in Table 8 (the cells with a cray background indicate relevant functions).

Table 8. Preservation Planning Functional Entity

Name	Description	Relevance to the research
Monitor Designated Community	Interacts with Archive Consumers and Producers to track changes in their service requirements and available product technologies.	Allows for collecting information about users to gain a better understanding of their knowledge and provide better access to archived digital knowledge.
Monitor Technology	Is responsible for tracking emerging digital technologies, information standards, and computing platforms (i.e., hardware and software) to identify technologies which could cause obsolescence in the Archive's computing environment and prevent access to some of the Archive's current holdings.	Allows for tracking the technology parameters used by the Designated Community.

This entity contains functions for monitoring the archive environment, providing recommendations, and ensuring that the information stored in the archive remains understandable by the Designated Community. Therefore, the main research question is about the designated community:

What method could be used to support profiling of the designated community?

6. Access Functional Entity

According to OAIS, the Access Functional Entity provides the services and functions that support Consumers in determining the existence, description, location, and availability of information stored in the OAIS, and allowing Consumers to request and receive information products (CCSDS 2012, pp 45) in Figure 15 (CCSDS 2012, pp 59).

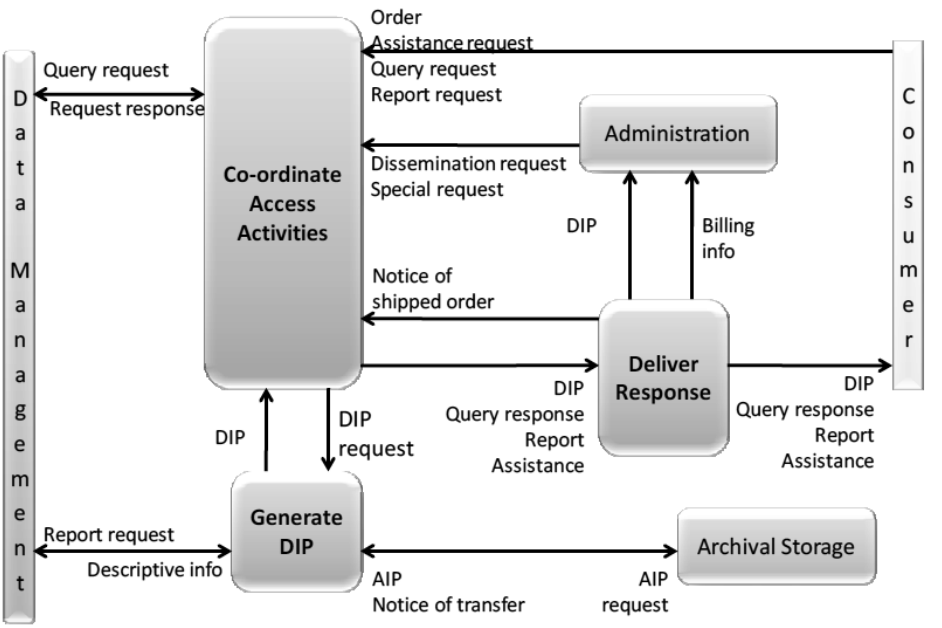


Figure 15. Access Functional Entity (CCSDS 2012, pp 59)

According to OAIS, the functions can be summarised as seen in Table 9 (the cells with a grey background indicate relevant functions).

Table 9. Access Functional Entity

Name	Description	Relevance to the research
Coordinate Access Activities	Provides one or more interfaces to the information holdings of the Archive.	Allows for providing better multifaceted access (which harmonises the knowledge bases of users) to archived digital knowledge.
Generate DIP	Accepts a dissemination request, retrieves the AIP from Archival Storage, and moves a copy of the data to a temporary storage area for further processing.	Does not include anything knowledge-specific and therefore is not observed in the research. The DIP can be presented in a number of ways (e.g. SOAP service, Markup Language (XML file), PDF file, Image (PNG), BASE64, etc.), which may not be related to any knowledge-specific properties and therefore is not a part of this research.
Deliver Response	Handles both on-line and off-line deliveries of responses (DIPs, query responses, reports, and assistance) to Consumers.	Does not include anything knowledge-specific and therefore is not observed in the research.

This entity contains functions for identifying, requesting, and accessing the archived content. Therefore, the main research question for this entity is related to the novel archival access to digital knowledge:

How can faceted browsing support the archival access to digital knowledge?

1.4. Research method

The research method of this work combines conceptual theoretical analysis of the main issues related to digital preservation of knowledge and several practical experimental projects.[3]

The research starts with a state-of-the-art analysis. Significant research projects related to this thesis will be identified and their results analysed. The overview will be presented as separate project cards which list all relevant characteristics of projects (see page 144).

All relevant core entities of the OAIS model will be identified for this thesis. The main research questions will be formulated and organised based on the selected entities – each research question will point to one or more OAIS functional entities (see page 30).

This will set the scene for the following research steps:

1. Identifying a proper workflow for an archival ingest based on Estonian requirements for archiving.
   A simple, but effective workflow must be in place in order to enrich the information during the submission. The designed workflow will be flexible and universal enough to cover even hybrid (both digital and paper) material archiving. The workflow will be introduced and tested in the paper I.

2. Updating the submission information package (SIP) specification of the National Archives of Estonia to support digital knowledge archiving.
   The proposed specification must be compatible with the requirements and metadata elements set by the archives, but on the other hand, it should be flexible and allow of incorporating new content-specific descriptions for digital records to support the raising of digital information to the next (digital knowledge) level. The specification will be introduced and explained in the paper I and on the dedicated web page (http://www.arhiiv.ee/en/universal-archiving-module) of the National Archive of Estonia.

3. Delivering a suitable pre-ingest tool which supports fulfilling the previously mentioned workflow and SIP specification.
   In order to test the theoretical work around the designed workflow and SIP specification in practice, the pre-ingest tool (Universal Archiving Module - UAM) of the National Archives of Estonia should be updated.

4. Conducting a pilot archival transfer project to test previous work.
   The pilot archival transfer will consist of archiving the records of the Office of Minister for Population and Ethnic Affairs of Estonia by using the UAM tool, the SIP specification and designed workflow. The pilot project will be described in the article I in more detail.

[3] The research tries to foster greater collaboration across disciplines by taking advantage of the synthesis of both traditional archival research and information technology methods.

5. Identifying, collecting, and investigating European best practices for archiving.
   Getting up-do-date information about the best practices in other European countries will allow making adjustments and improvements in the previous work. Therefore, the author of this thesis will also participate in the task of delivering the report D3.1 E-ARK Report on Available Best Practices.

6. Developing a new approach for identifying the designated user community.
   Although the OAIS model describes the need for identifying the designated community as one of the key tasks, there is no good method for that purpose available. Therefore, the topic will be analysed more closely and a new method will be proposed based on the research. This research step will be described in the paper II.

7. Developing new user-centred access functionalities for the archival access portal.
   This research step concentrates on end-user knowledge base and novel access techniques. An update of implementing predefined facets will be made to the archival access portal to support multifaceted access to the content. These facets will enable users to adapt applications faster because they can give results with less user effort. This research step will be described in the paper III.

8. Developing a solution for supporting user knowledge transformation to archival knowledge.
   The users of archives may have new knowledge about the records held in the archives. Their knowledge may help to cover important gaps in archived knowledge. Therefore, this task will explore methods to capture user knowledge and transform it to archival knowledge. This research step will be described in the paper IV.

9. Developing a common (international) SIP specification for archival knowledge ingest.
   Many different SIP specifications are used all over the world and more harmonisation is needed in order to achieve interoperability in knowledge ingest. Instead of creating a totally new specification or insisting organisations to use some available standard, a harmonised version of the SIP format will be provided by E-ARK. Organisations can continue to utilise familiar tools as only some minor updates has to be made to support the specification. Therefore, the author of this thesis will also participate in the task of delivering the SIP specification for the E-ARK project.

The research's theoretical and practical implementations will be developed and executed in cooperation with the National Archives of Estonia and the E-ARK project.

2. REALISATION, RESULTS AND DISCUSSION

The research was carried out by following the previously described research method.

Four scientific papers were published in journals and/or conference proceedings during the research. The complete list of papers can be found in the list of original publications on page 8 and full texts of each publication can be found in the publication section starting from page 77 of this dissertation. Both the articles and research questions are tightly related to OAIS. The relations with OAIS entities can be seen from Figure 16:

- The paper I and the question "How can the ingest function be broadened to help prepare records for archiving already in the pre-ingest stage" are related to Ingest and Administration.
- The paper II and the question "What method could be used to support the profiling of a Designated Community?" are related to Preservation Planning.
- The paper III and the question "How can faceted browsing support the archival access to digital knowledge?" are related to Access.
- The paper IV and the question "How can users share their knowledge with archives?" are related to Data Management, Access and Archival Storage entities.

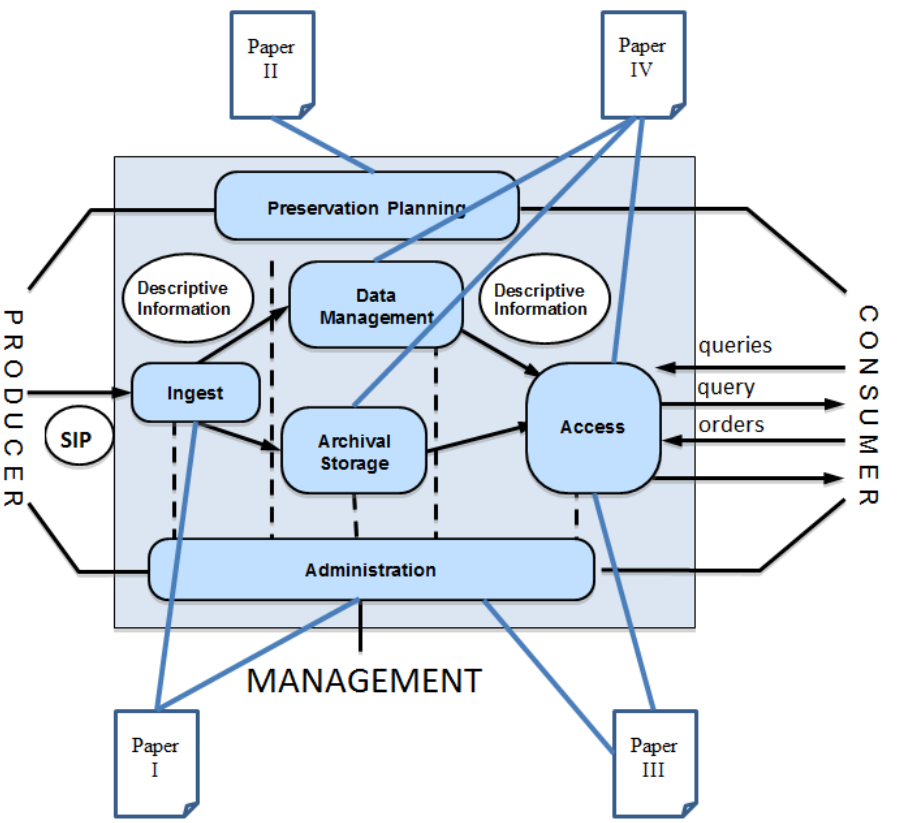


Figure 16. The relation between research papers and OAIS

The current chapter will explicate the context of research in more detail by discussing the papers in the context of research questions.

2.1. How can the ingest function be broadened to help prepare records for archiving already in the pre-ingest stage

Paper I (Digital Preservation of Knowledge in the Public Sector: A Pre-Ingest Tool) explored the questions how can the ingest function be broadened to help prepare records for archiving already in the pre-ingest stage by adding detailed metadata about the structure, context, and relationships to transform the information to digitally codified knowledge?

The paper described a pilot project of using the updated submission information package specification and the relevant pre-ingest tool in a real practice. A simple workflow for knowledge ingest was introduced and followed (Figure 17).

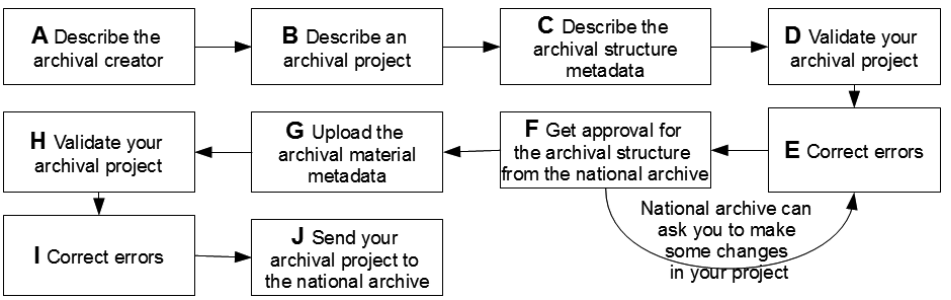


Figure 17. Workflow for the SIP preparation

According to the workflow, the preparation and transfer of archival records can be divided into ten phases (as described in the paper I), starting from the description of the archival creator and ending with the controlled transfer of records and their metadata to the archive.

The workflow helped to create a clear vision and to track the progress of the archival project at any time.

The research showed that it is very important to pay special attention to the pre-ingest stage by pointing out the fact that updates to the SIP should be made as early as possible – add metadata about important characteristics already at pre-ingest or during the preparation process. It is crucial to build a classification schema, add descriptive metadata, and describe also the digital content itself (e.g. automatically identify and characterize the computer files). This all led to the fundamental recognition that even more important than the ingest function is the pre-ingest which is not represented in the list of original entities of OAIS model.

The pre-ingest function should also include a multi-level validation system, which can help to ensure that the descriptions and the internal structure of records (structure of archives) is compiled in accordance with the requirements of the archive in a less time- and labour-consuming way.

This can be achieved by using a pre-ingest tool like UAM (Figure 18). The tool can be configured to use some specific electronic records management system (ERMS) in order to be able to reuse the descriptions that are available in ERM systems when describing the archival items.

Some large agencies may simultaneously use several systems for managing records (e.g. ERMS X1, ERMS Xn) and these systems may also

- have different metadata sets (M_{X1} and M_{Xn})
- contain a variety of computer file formats (F_{X1} and F_{Xn})
- have ability to dispose of records in a specific format (D_{X1} and D_{Xn})

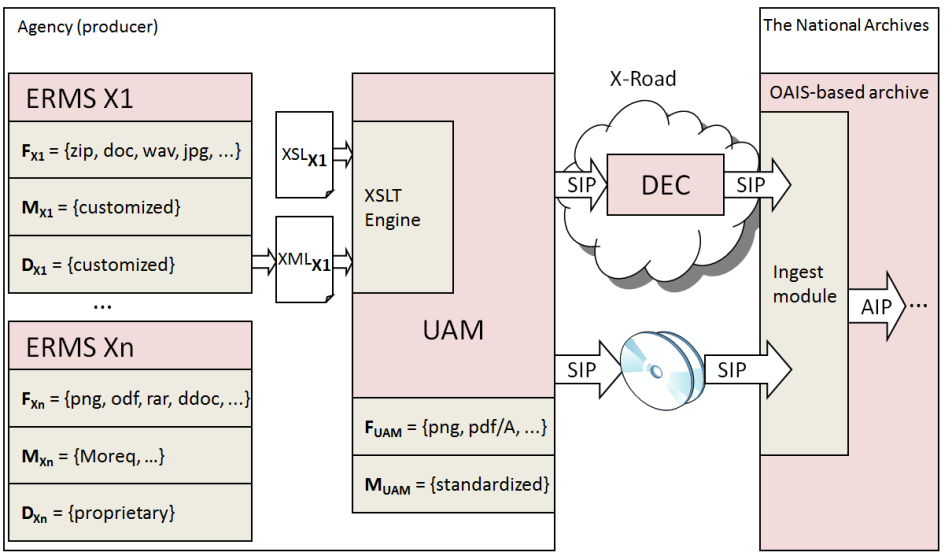


Figure 18. Technical view to pre-ingest activities

These formats do not always correspond with the metadata set and file formats specified in the UAM and therefore an XML transformation (XSLT engine) should be used to transform the exported content and metadata into a suitable format. For example, it is possible to transform an element named <DocName> (from ERMS metadata) to an element <Title> in UAM by mapping the ERMS output to the UAM's input scheme with an XML transformation.

The UAM also provided the following functionalities to the agency's archivist:

- building a classification schema (both manually and automatically from XML files),
- adding descriptive metadata (both for paper based and digital records, manually or automatically),

- describing digital content (automatically identifying and characterizing the computer files),
- validating (automated control against rules and requirements set by several laws and guidelines) the SIP.

The SIP was created and transferred to the National Archives of Estonia by using the same tool. The submission was double checked during the ingest and sent to the archival repository afterwards. The validation in ingest module confirmed the quality of the SIP and allowed to evaluate the success of the preparation and transfer.

The paper concluded with the statement that it is important to act proactively at the (pre)-ingest stage and pay very close attention to how the ingest process is being conducted, as this is the best stage for the transformation from information to knowledge since producers and submitters know their material best and can influence the quality of submissions most.

In the sense of OAIS, functions Negotiate Submission Agreement, Establish Standards and Policies, Audit Submission, Receive Submission and Quality Assurance in Ingest were covered.

2.2. What method could be used to support the profiling of a Designated Community?

This research question was studied in the article II (Digital Preservation and Knowledge in the Public Archives: For Whom).

One of the key requirements in digital preservation is that the archived digital knowledge should be independently understandable to a designated community, and therefore it is important to know who the main users of the archives are.

Paper II points out that the designated community can be managed mainly through the Monitor Designated Community (MDC) function in OAIS (Figure 19).

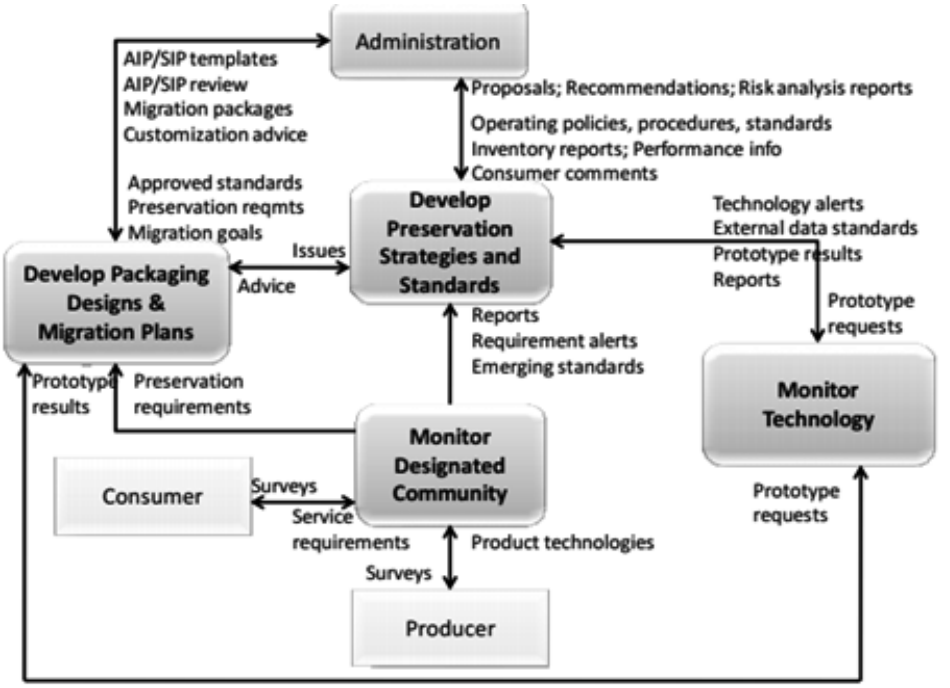


Figure 19. Monitor Designated Community function in OAIS (CCSDS 2012, pp 57)

As the OAIS model does not give any recommendation for the actual implementation of MDC then the paper went deeper in research and proposed the use of automated user observation on the web and user profiling by using a special model and a method to support the profiling process (Figure 20).

This proposed TRIANGULAR method supports obtaining more information about users, their knowledge bases and tracking the changes over time. These are also crucial requirements, as it is well known that the knowledge base of the designated community may change in time and that some important aspects may no longer be understandable then.

The proposed model is a good example of applying techniques from other disciplines in the digital preservation area. The model helps to identify the designated community in a relatively simple way and to update the profiles flexibly if the designated community changes over time.

The paper also introduced a universal scale (from 0 to 100) for evaluating profile results by acknowledging user engagement. For example, if users from the primary users group have searched the keyword 'maps' the most (e.g. 150 times), it will be valued as 100, and the keyword with the fewest searches (e.g. 3 times) will be valued as zero.[4] The zero value will be automatically excluded from the significant results list. All other values will be interpolated to

[4] If there is more than one minimum or maximum, they still will receive the same values.

transformation values from 1 to 100. Transformed values will have different weights depending on the user group (Table 10):

- the primary user group will have a weight of 75%,
- the secondary 50%,
- the tertiary 25%.

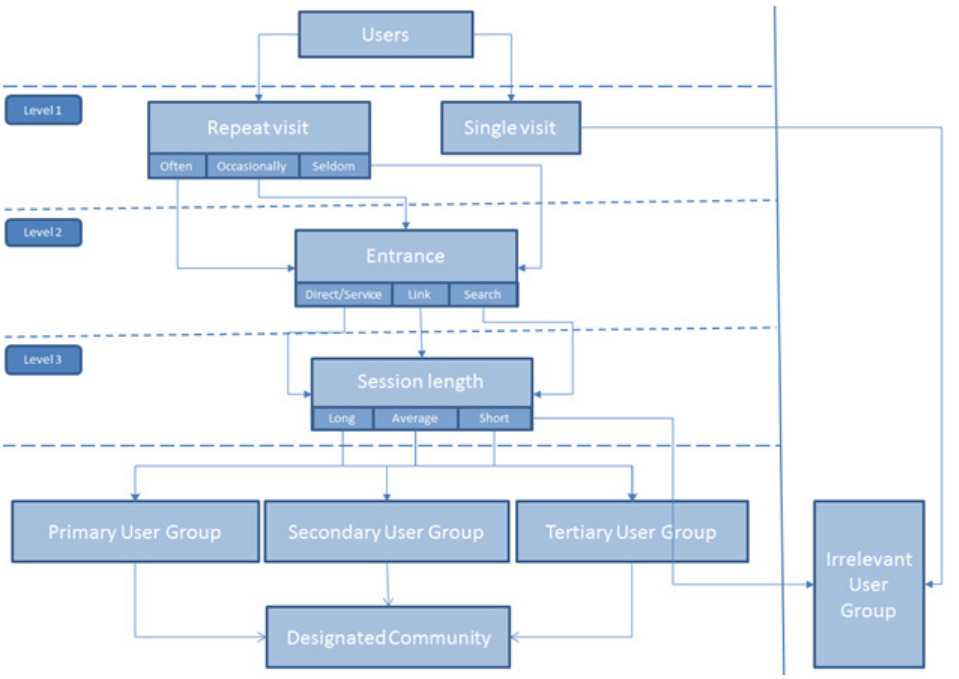


Figure 20. User profiling model

Table 10. Functions of the preservation planning functional entity

	Minimum I Quarter	I Quarter	II Quarter	III Quarter	Maximum IV Quarter
Primary		×	✓	✓	✓
Secondary		×	×	✓	✓
Tertiary		×	×	×	✓

The main technological parameters (e.g. browser, screen resolution, OS, plugins) of the systems used by the designated community can be also easily identified by following the TRIANGULAR model.

The article points out that it is reasonable to use a combination of user observation methods, because the web analytics can be influenced by visitor's settings (e.g. the user has disabled JavaScript or blocked cookies).

The second important recommendation is that in interpreting results we should not make assumptions that are too general. For example, if notice that visitors do not use breadcrumbs on the top of our web page then we cannot be certain in the reason without further investigation (for example, the bread-crumbs may be in a place which is too hidden, the font is too small, etc.).

The paper took also into account the previous research (especially the research on formalising the intelligibility using modules and dependencies) carried out in the CASPAR project when developing the new approach for user profiling.

In the sense of OAIS model, the paper II covers both functions Monitor Designated Community and Monitor Technology of the Preservation Planning Functional Entity and is partially related to the Establish Standards and Policies function as well as one of the missions of this entity is establishing a policy for designated community management.

2.3. How can faceted browsing support the archival access to digital knowledge?

This research question was studied in the paper III (Toward Common Ontologies of Facets of the Archival Access Portal).

The paper III reflects the possibility of providing objective access to the designated community even if the access is influenced by users' knowledge and is not homogeneous. The paper points out that user interface of access system should be flexible and convenient, and it should take into account the knowledge base of the designated community.

The paper introduces one yielding approach for providing access to archived content by using advantages of a multi-faceted archival access portal of the National Archives of Estonia (NAE). The paper points out that applying faceted browsing capability to the archival access system can be a challenge to the memory institution, as it may require a lot of additional activities (e.g. taxonomies) to be carried out before it can be implemented.

Therefore, the paper also explains the development and implementation process for five taxonomies and classifications which were the basis for the facets.

It also demonstrates that it is important to have and use common guidelines and best practices for encoding different metadata schemas and implementing specific organized interpretation of knowledge structures (e.g. a thesaurus).

Developing faceted browsing support to the archival access was practically demonstrated by updating the new archival access portal of the National Archives of Estonia (Figure 21).

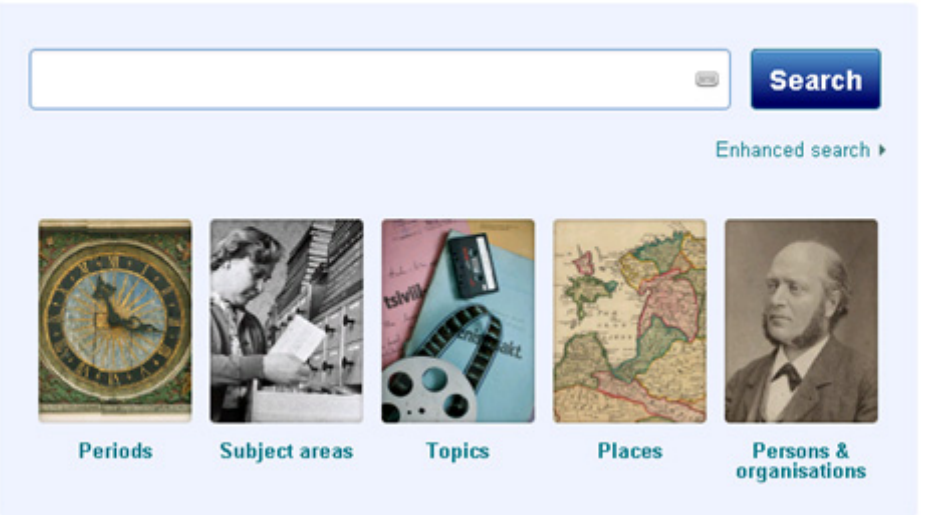


Figure 21. Simple search with five facets

Those defined facets (periods, subject areas, topics, places, persons and organisations) on the first page of the simple search can act as a starting point for browsing if the user does not know what to type in the search form.

The user does not have to type symbols in the search box, because the search and browsing can be started by clicking on one of the images. This allows the archival access portal to be used as a selection-based search system in which the user can invoke a search query by using only the mouse or by tapping the screen of the tablet.

Each click on the facet name will narrow the search by providing a new selection of available facets. Only these facets which can currently be used will be displayed to the user (Figure 22. Guided navigation).

Figure 22. Guided navigation in the access portal

When the facet which has sub-categories is clicked the sub-categories with the number of linked values will be revealed and the user can narrow the search even more. Such iterative search helps to see what kind of content is currently left to browse and therefore no empty results will be provided to the user.

As such, this approach may lower search barriers, especially when the user has not mastered computer use or his/hers knowledge base significantly differs from the designated community's knowledge base and skills.

The paper points out that creating the relationships between the archival content and appropriate taxonomies will continue. This is a challenging task as there are many situations when the human experts could only decide what the exact value of a facet should be and to what content it should be linked. Sometimes even reading the content would be required to identify the right classification.

In the sense of OAIS model the paper III covers the Coordinate Access Activities (CAA)[5] function in the Access Functional Entity as it investigates the multifaceted access interface to the information holdings of the archive.

2.4. How can users share their knowledge with archives?

This research question was studied in the paper IV (Transforming users' knowledge into archival knowledge).

It is most likely that the archival creators know the information to be archived best, so they can also enrich it most effectively, but there are many archives which have already ingested digital information which could be enriched by other interested parties. Involving crowd knowledge would be very useful to archives in many sense. On one hand external users can raise the quality of archival holdings (for example describe the people, places and activities depicted on photos), but on the other hand save the time of archivists.

In such cases we can take the advantage of the Data Management Functional Entity – DMFE (in cooperation with the Archival Storage Functional Entity – ASFE) which allows for enriching the digital information held in the digital repository. More precisely, DMFE includes a Receive Database Update function, which allows for adding, modifying or deleting information in the Data Management's persistent storage (Figure 23).

According to the OAIS model, the Ingest Functional Entity is expected to coordinate the updates between DMFE and ASFE, but this may involve several complications in real life situation:

- OAIS ingest is designed for the one-collection-at-a-time principle of receiving, but the enrichment process can involve information from multiple fonds/collections and submissions;

[5] The CAA could also act as a gateway for the enrichment workflow as proposed in the article IV.

- even small updates will activate most of the ingest workflow steps, which may not be an optimal use of resources;
- quality control can differ: the archivists at the archival institutions are generally not required to check the veracity of Submission Information Package (SIP) descriptions (that is usually the responsibility of the archivists at the producer side), but in the case of crowdsourcing, the updated descriptions proposed by users have to be manually double-checked and confirmed before being accepted as part of official archival descriptions.

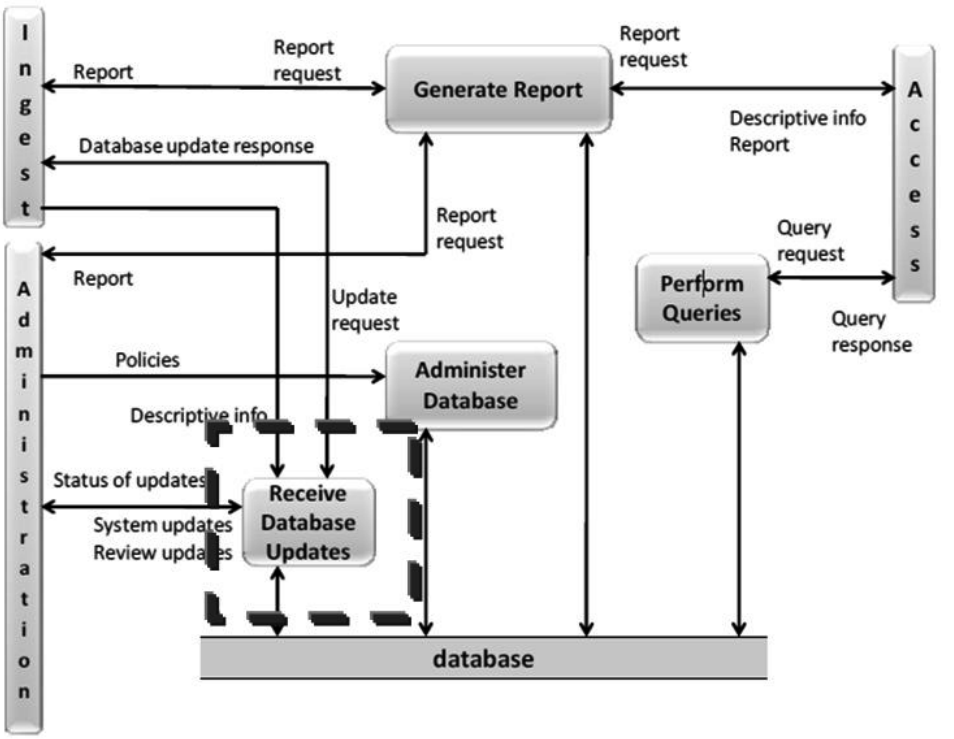


Figure 23. Data Management Functional Entity (CCSDS 2012, pp 53)

Therefore, involving the ingest functional entity may not be the most optimal way for updating descriptions via crowdsourcing. These complications are not critical for every use case, but for the needs of NAE it made sense to look for a more efficient solution.

The NAE devised a plan for that purpose. The plan included the process of

- generating Persistent Uniform Resource Identifiers for archival holdings
- defining the archived digital information as resources (RDF)
- making archival holdings public (including taxonomies)
- developing selecting functionality for sending information to specialised crowdsourcing tools

- developing a tool for presenting the information sent to crowdsourcing
- developing functionality for receiving crowdsourced information

The research realised that by creating a direct connection between the Data Management Functional Entity (more precisely the Receive Database Updates function) and Access Functional Entity in the OAIS model we could serve the crowdsourcing interaction with external users in a more effective and intuitive way (Figure 24).

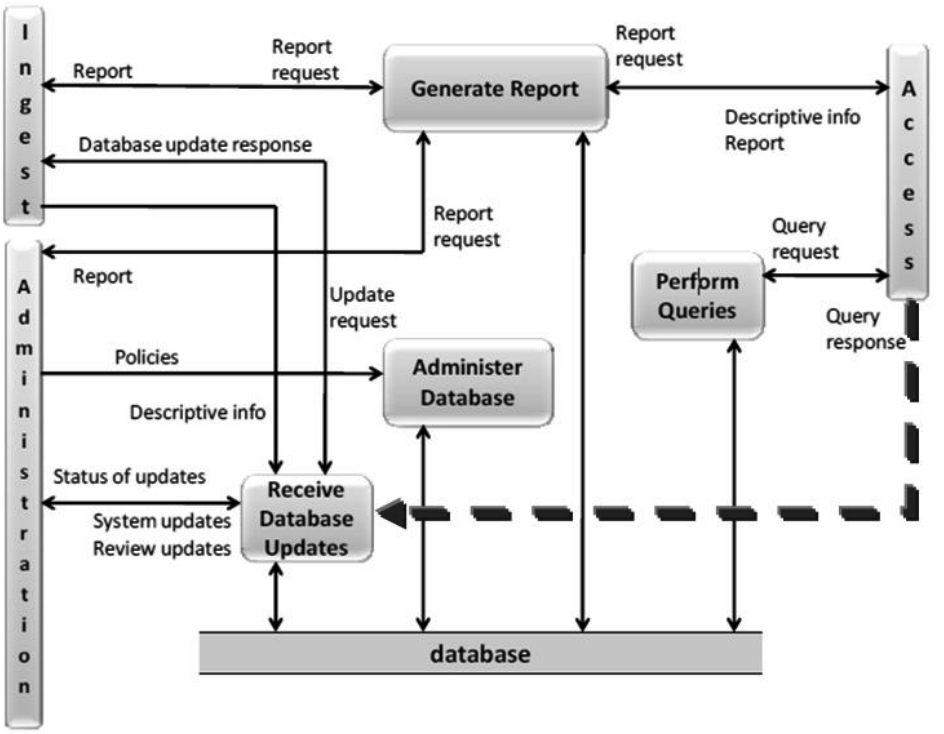


Figure 24. A direct connection between the Data Management and Access entities (CCSDS 2012, pp 53)

The paper concluded that the access entity should act as a gateway in interaction with external users. The archivists select and send the information to crowdsourcing in the access entity and also receive the enriched information back through the same channel. Enriched information will then be sent to the Receive Database Updates service in DMFE for further actions. This complemented approach for database updates is more intuitive and requires less resources than using the full OAIS Ingest workflow.

The practical outcome of this work was an application interface for getting information out from the archival catalogue for enrichment through crowdsourcing and getting the results back later.

3. CONCLUSIONS

Digital preservation of knowledge is a very broad and complex research area. Many aspects are still open for research. According to the literature, the accessibility and usability of digital information have been more investigated than the comprehensibility of important digital information over time. Therefore, this doctoral thesis examines the continuous comprehensibility and the meaningful use of digital objects by taking into account the designated users' knowledge and OAIS model. Some new mechanisms will be provided to complement the entities of the OAIS model through practical experience and interdisciplinary actions to support the digital preservation of knowledge. Furthermore, this work explores the OAIS model from real practical experience and gives a broader view of knowledge than that originally defined in OAIS. It brings out the issues in OAIS regarding digital preservation of knowledge and seeks practical solutions based on OAIS' functional entities. More precisely, it seeks answers to the following questions:

1. How can the ingest function be broadened to help prepare records for archiving already in the pre-ingest stage by adding detailed metadata about the structure, context, and relationships to transform the information to digitally codified knowledge?

   The research showed that it is very important to pay special attention to the Pre-Ingest stage. A practical experiment was conducted to prepare and transfer a well-defined (well-described attributes of the content and context to justify the quality criteria for knowledge) SIP to the archives and the idea proposed in the previous question was proven.

   This research question was studied in Paper I (Digital Preservation of Knowledge in the Public Sector: A Pre-Ingest Tool).

   The practical outcome was an update of the pre-ingest tool universal archiving module (UAM) to support the theoretical work (the SIP specification and the pre-ingest, ingest workflow).

   The SIP specifications can be seen in appendices:

   - Appendix B. Local Submission Information Package Specification
   - Appendix C. International Submission Information Package Specification

2. What method could be used to support the profiling of a Designated Community?

   This research question was studied in Paper II (Digital Preservation and Knowledge in the Public Archives: For Whom).

   Archived digital knowledge should be independently understandable to a Designated Community, and therefore it is important to know who the main users of the archives are. The solution can be facilitated using the TRIANGULAR method. This also supports obtaining more information

about their knowledge bases and tracking the changes over time. These are also crucial requirements, as it is well known that the knowledge base of the designated community may change in time and that some important aspects may no longer be understandable then.

The practical outcome was an example implementation based on the open source web analytics platform Piwik.

3. How can faceted browsing support the archival access to digital knowledge?

   This research question was studied in Paper III (Toward Common Ontologies of Facets of the Archival Access Portal).

   The user interface of access system should be flexible and convenient, and it should take into account the knowledge base of the designated community. Applying faceted browsing capabilities to the archival access system can be a challenge to the memory institution, as it may require a lot of additional activities to be carried out before it can be implemented. Developing faceted browsing support to the archival access was practically demonstrated by updating the archival access portal of the National Archives of Estonia.

   The practical outcome was the update of the archival access portal to support faceted browsing.

4. How can users share their knowledge with archives?

   This research question was studied in Paper IV (Transforming users' knowledge into archival knowledge).

   It is most likely that the archival creators know best the information to be archived, so they can also enrich it most effectively, but there are many archives which have already ingested digital information which could be enriched by other interested parties. This was also the case for the National Archives of Estonia. In order to prepare the archived information (including archival taxonomies) for such crowdsourcing activities, they were first defined as resources and then opened to the public in a machine-readable way. This allows not only crowdsourcing, but also application of other innovative technologies (like pattern-based data mining) which have been actively used only in other disciplines thus far.

   The practical outcome was providing a means for the transfer of users' knowledge into archival knowledge.

Further research

It is clear that digital preservation of knowledge is not a state or a place where someone can arrive and stay. It is a continuous journey, a combination of constant monitoring and change management. Therefore, current dissertation does not aim to cover all possible aspects (a lot of them are already covered by a number of projects, as indicated on the project overview cards) of the digital preservation of knowledge, but it gives an insight to some of the current relevant issues that affect the area.

This work delivered some answers regarding the digital preservation of knowledge, but there are still some tasks that need to be researched by further studies:

- Investigate the applicability of the SIP specification in other domains as business and personal archiving sector is also searching for suitable specification for knowledge ingest.
- Study the effects of user profiling in real practice. It will require of the development of prototypes to use in testing of the suggested approach.
- Extend access facets by building links with other taxonomies. This will prevent archived digital resources from becoming "information islands" as the knowledge is dynamic in essence and can acquire new relations with other repositories if ontologies are shared.
- Investigate the benefits of users knowledge transfer in measurable numbers. By measuring the value of crowdsourcing can help to encourage the memory institutions to open up their fonds for enrichment.
- Develop a standardised means of communication between catalogue/ repository and crowdsourcing systems.

SUMMARY IN ESTONIAN

Digitaalse teadmuse arhiveerimine – teoreetilis-praktiline uurimistöö Rahvusarhiivi näitel

International Data Corporationi (IDC) poolt 2013. aastal läbi viidud uuringu järgi digiteeriti või sündis juba digitaalsena ligikaudu 4,4 ZB (zettabaiti) informatsiooni (EMC Digital Universe, IDC 2014), mis on hinnanguliselt 1 triljon tavapärast DVD-R-plaati (*Digital Versatile Disc Recordable*). Ühest küljest tekitab selline suur hulk infot kartust, et vajalikku ei leita enam mõistliku ressursikuluga üles, kuid teisest küljest on põhjust tunda muret ka olulise informatsiooni säilimise pärast. 2009. aastal viidi projekti "Planets" raames arhiivides, raamatukogudes ja muudes avaliku sektori asutustes läbi uuring digitaalse säilitamise alase teadlikkuse väljaselgitamiseks, mille tulemuste järgi 93% küsitletutest teadvustab digitaalse säilitamise väljakutseid, kuid ainult 25% neist omab või kavandab vastavat digitaalarhiivinduslikku lahendust (Tessella, British Library 2009).Selline kardinaalne erinevus on tingitud tõenäoliselt eelkõige asjaolust, et digitaalsel informatsioonil on tegelikult mitmeid varjatud külgi, millega peab alati arvestama ja millele lahenduste leidmine on ressursimahukas. Kuna digitaalne informatsioon on oma olemuselt väga keeruline ja habras, siis tuleb teha pingutusi, et garanteerida tuleviku ühiskonna jaoks vajaliku teabe olemasolu/kättesaadavus, kasutatavus ja mõistetavus.

Olemasolu/kättesaadavus

Digitaalses maailmas kasutatavad tehnoloogiad vajavad tihtipeale spetsiaalseid töö- ja hoiutingimusi ning käsitlemisoskusi. Võrreldes n.ö paberipõhise maailmaga võib tunduvalt hõlpsamini juhtuda, et ebaõigete töövõtete või -tingimuste tõttu leiab aset midagi pöördumatult kahjulikku kas andmekandja ja/või seadmega. Näiteks võib kompaktplaadi (CD) pealmisele poolele (kogemata) tekkinud 8 mikroni (0,008 mm) sügavune kriim rikkuda selle andmekandja, sest CD-plaadile kirjutatud andmed on kaetud vaid umbes 6–7 mikroni paksusega lakikihiga (Mueller 2004). Arvuti kõvaketta kahjustamiseks piisab aga näiteks 10 V suuruse pinge rakendamisest (IT Business Edge 2003). Sellist pinget on inimestel äärmiselt lihtne tekitada, sest tavapärane, nt juuste silitamisest tekkiv staatilise elektri laeng on vähemalt 1500 V, aga kui laengu vabanemisel on näha sädet, siis on pinge juba vähemalt 8000 V (Andrews 2009). Kui vaadelda samu näiteid pabermaailma kontekstis, siis suure tõenäosusega on võimalik konserveerijatel eelpool kirjeldatud kriimustus edukalt parandada ja mainitud 10 voldise elektrilaengu puhul ei pea paberdokumendiga üldse midagi ette võtma.

Kuna digitaalsel ainesel peab olema alati siiski ka mingisugune füüsiline kuju (nt magnetketas), siis on ühtviisi ohtlikud kõik füüsilised õnnetused ja loodushädad, mis ka analoogmaailmas probleeme põhjustavad (üleujutused, tulekahjud, tormid jms).

Nii nagu aja jooksul muutub tint paberil halvemini loetavaks või paber hapramaks, toimuvad muutused ka digitaalsete andmekandjatega. Näiteks CD-plaatide elueaks pakutakse erineva kirjanduse järgi ideaaltingimuste korral 5–59 aastat (Rothenberg 1999). Kuna laserplaadid ei ole paraku nii kaua veel käibel olnud (tänapäevased laserplaadid tulid kasutusse 1980ndatel), et saaks nii pikka eluiga kui 59 aastat reaalse kasutuskogemuse põhjal lubada, siis tuleb arvestada, et tegu on siiski arvutuslike või laborikatsetega.

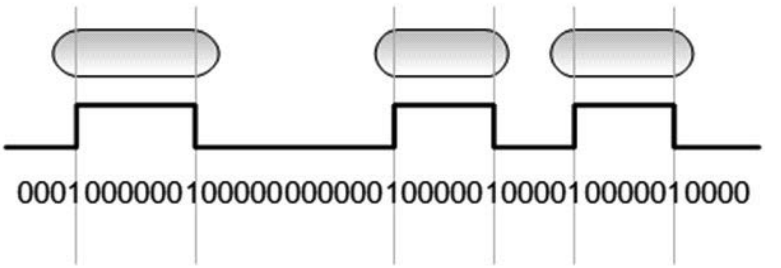
Hoolimata sellest, et osa andmekandjaid on spetsiaaltingimustes hoides füüsiliselt püsivad, toimub siiski tehnoloogia iganemine. Selle taga on asjaolu, et tehnoloogia uuenemise tsükkel on lühike, 3–5 aastat (Thomaz 2006). Nii leidub paljudel inimestel veel kümmekond aastat tagasi rohkesti kasutatud 3.5" magnetkettaid, kuid seadmeid, millega neile talletatud andmed kätte saaks, on vähestel. Kuna saadaval on andmekandjad, mis mahutavad palju rohkem andmeid, siis varasema tehnoloogiaga seotut ei ole tootjatel enam otstarbekas luua või arendada. Seadmete ja andmekandjate iganemine ning käibelt kadumine on probleem, millele ei ole üheselt head lahendust peale pideva tehnoloogiaseire ning riistvara ja andmekandjate värskendamise (nt CDde ümbersalvestamine DVDdele). Lisaks on soovitatav juba andmete salvestamisel valida võimalikult püsiv andmekandja. Samas peab arvestama, et tehnoloogia arenedes tekib ikka uusi, paremaid (suurema mahutavusega, püsivamaid jne) andmekandjaid ja siis tuleb jälle õigel hetkel vanadele kandjatele talletatud sisu uutele üle viia.

Radikaalsema lähenemisena on kasutusel ka n.ö arvutimuuseumi loomine, st et koos andmekandjatega säilitatakse ja hoitakse kasutuskorras nende jaoks vajalik originaalne riist- ja tarkvara. Antud lahendus on äärmiselt kulurikas, kuna kogutavate süsteemide hulk kasvab väga kiiresti. Hüppeliselt kasvavad kulud eriti juhul, kui tahetakse välistada seadmete riketest tulenevaid kahjusid seadmete dubleerimisega. Eksperdid on ühesel arvamusel, et see lahendus ei ole jätkusuutlik (Borghoff 2006).

Digitaalse informatsiooni kättesaadavus on mäluasutustes üsna hästi tagatud, kuna osatakse vastavaid riske teadvustada. Rahvusarhiiv peab mõistlikuks maandada kättesaadavusega seotud riske andmete dubleerimisega, varundades teavet mitmes geograafiliselt erinevas asupaigas, ja valides erinevad salvestus-meediumid, et välistada nt ühe seadme tüüpvea kordumist dubleeritud andmetes (Rahvusarhiiv 2008a). Lisaks sellele on tähtsal kohal ka riistvara ja salvestus-meediumite nõuetekohane hoidmine ning käsitlemine.

Kasutatavus/esitatavus

Järgmise kitsaskohana saab välja tuua asjaolu, et digitaalse informatsiooni füüsiline bitiline esitus ei defineeri üheselt selle loogilist esitlust. Näiteks kui vaadata piisavalt võimsa mikroskoobiga kompaktplaati, siis võib näha laseriga plaadile lõigatud süvendeid (1. joonis). Bittide salvestamisel on kasutatud algoritmi, mille järgi alustatakse ja lõpetatakse plaadi pigmendikihi söövitamine biti oleku 1 korral. See tähendab, et plaadil olev tühimik ei vasta otseselt ainult ühele bitile ega selle olekule. Seega peab olema teada lisainfo algoritmi

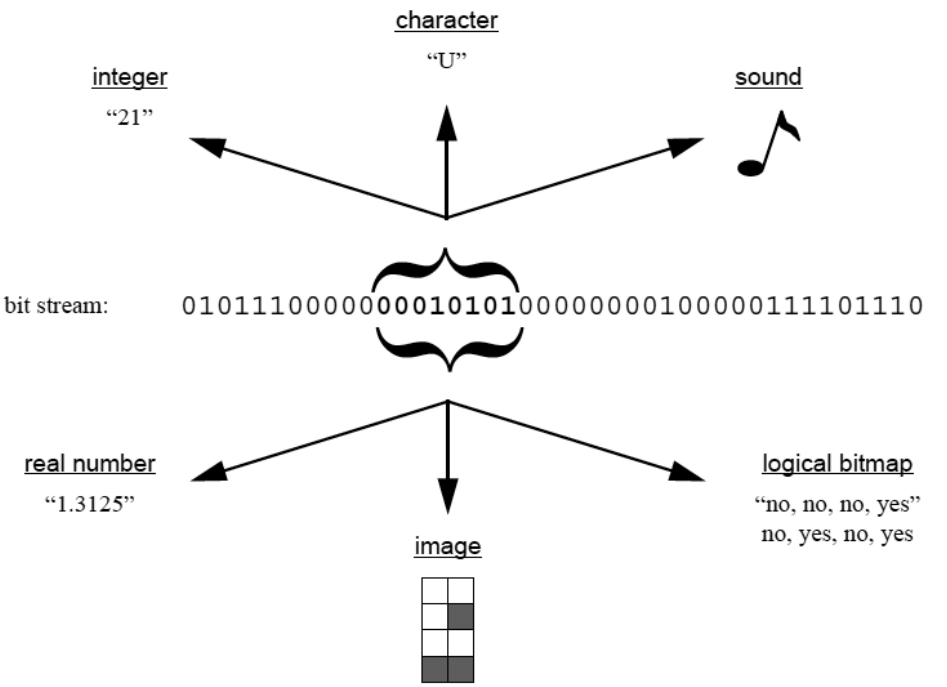


Joonis 1. Kompaktplaadile salvestatud bitid

kohta, mis annab meile selgitust, kuidas andmeid salvestada ja, veelgi olulisem, kuidas andmeid välja lugeda. Antud juhul peab olema teada, et ainult süvendi ääre avastamisel lisatakse bitijadasse bitt väärtusega 1, kõikidel teistel juhtudel on väärtusteks nullid.

Sisendseadmetega suhtlemiseks tuleb kasutada draivereid ja mällu loetud bittide tõlgendamiseks operatsioonisüsteeme ning rakendustarkvarasid.

Andmekandjalt mällu loetud bitijadasid saab aga samamoodi mitmesuguse loogika järgi edasi tõlgendada, mistõttu tekib vastavalt interpretatsioonile ühe bitijada kohta mitu õiget, kuid kujult erinevat esitlust. Tuleb teada, kuidas bitte grupeerida ning millist kooditabelit rakendada. Järgneval joonisel (2. joonis) on näha, et üht mällu loetud bitijada lõiku on võimalik mitmeti tõlgendada.



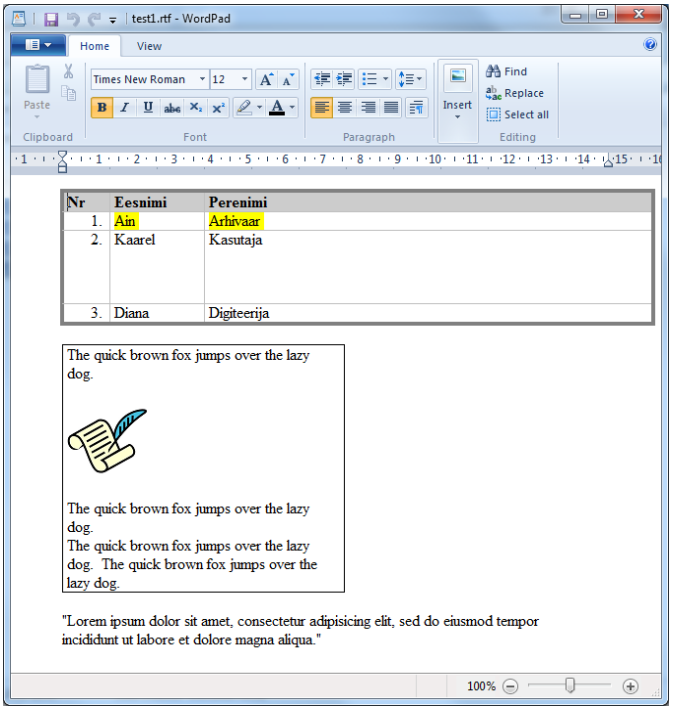
Joonis 2. Bitijada tõlgendamine erinevate esitlustena (Rothenberg 1999, pp 7)

Siinkohal muutub oluliseks juba failivorming, st kodeering on reaalselt seotud nii andmetüübi kui rakendusprogrammiga. Andmetüüpe võib eksisteerida väga palju – tekst, kaart, video, heli, graafik, joonis, valem, foto, noot jne. Iga tüübiga võib olla omakorda seotud palju kodeeringuid/vorminguid. Pidevalt tuleb kasutusele uusi vorminguid ja samal ajal on ka neid, mille aktiivne arendamine/toetamine lõpetatakse. See omakorda tähendab, et järgmise riistvara ja/või operatsioonisüsteemi jaoks ei pruugi enam olla sobivat tarkvara. Probleemi lahendamiseks on kaks peamist suunda: emuleerimine ja migreerimine.

Emuleerimise korral luuakse virtuaalne keskkond, mis matkib vana süsteemi ja võimaldab niiviisi vanu tarkvarasid ja faile kasutada. Heikki Vallaste defineerib, et emulaator on seade või programm, mis imiteerib teise seadme või programmi tööd, ning lisab, et tänapäeval võib termin "emulaator" käia nii riist- kui tarkvara või ka nende kombinatsiooni kohta (Vallaste 2012). Selliste lahenduste loomine on väga keerukas ja kulukas, kuna tuleb arvestada paljude variatsioonidega. Üks enamlubavamaid projekte antud vallas on "Keeping Emulation Environments Portable" (KEEP). Loodud on emuleerimise raamistiku tuum, mis võimaldab samas virtuaalmasina toel ka väliste emulaatorite lisamist (KEEP 2015). Emuleerimise oluliseks tugevuseks võib pidada asjaolu, et protsessi käigus ei muudeta esialgset bitijada, st objekt säilib autentsena. Paraku ei ole emuleerimise tulemuse korrektsuse hindamiseks teada üheselt aktsepteeritavat viisi peale visuaalse võrdluse. Seetõttu on paralleelselt levinud **migreerimise** taktika, kus failid konverteeritakse ümber moodsamasse, arhiivipüsivamasse vormingusse või uuendatakse nende versiooni. Uue vormingu valimisel arvestatakse selle kasutajaskonna suurust, lähtekoodi avatust, tarkvarade rohkust ja standardiseeritust (Rahvusarhiiv 2008b, lk 16). Migreerimise puhul tekib autentsuse tagamise probleem: kas migreeritud objekt on ikkagi seesama objekt. Lisaks ei aita migreerimine näiteks sellistel juhtudel, kui ei leita uut sobivat vormingut (nt arvutimängud). Seega peab tehnoloogia muutuste seiramine ja valikute kaardistamine olema mäluasutuste igapäevane tegevus, sest ühte kindlat viisi, mis sobiks igas olukorras digitaalse informatsiooni kasutatavuse tagamiseks, ei eksisteeri.

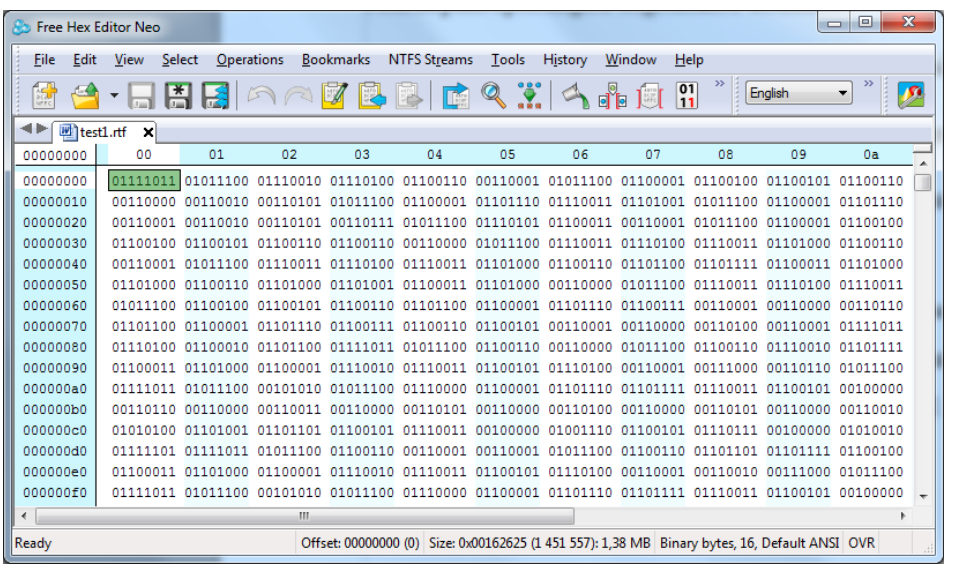
Mõistetavus/arusaadavus

Nii emuleerimine, migreerimine kui ka nende kombineerimine aitab säilitada informatsiooni kasutatavust, kuid ei pruugi tagada autentset mõistetavust, sest digitaalse teabe esituskuju ei ole jäigalt fikseeritud – esitus tekib igal korral alles tarkvara abil bittide konkreetse tõlgendamise kaudu. Näiteks tarkvaraga WordPad versioon 6.1 loodud tavalise dokumendi originaalilme on näha 3. joonisel.



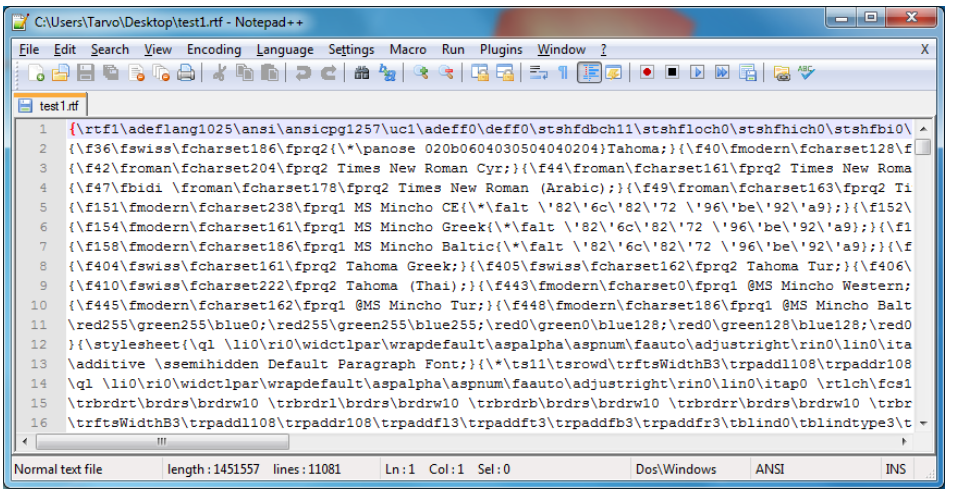
Joonis 3. WordPad versioon 6.1 vaade

Kui avada sama dokument Hex Editor Neo abil, siis on näha faili kahendkuju, mis on samuti õige tõlgendus, kuid tavakasutajale otstarbetu (4. joonis).



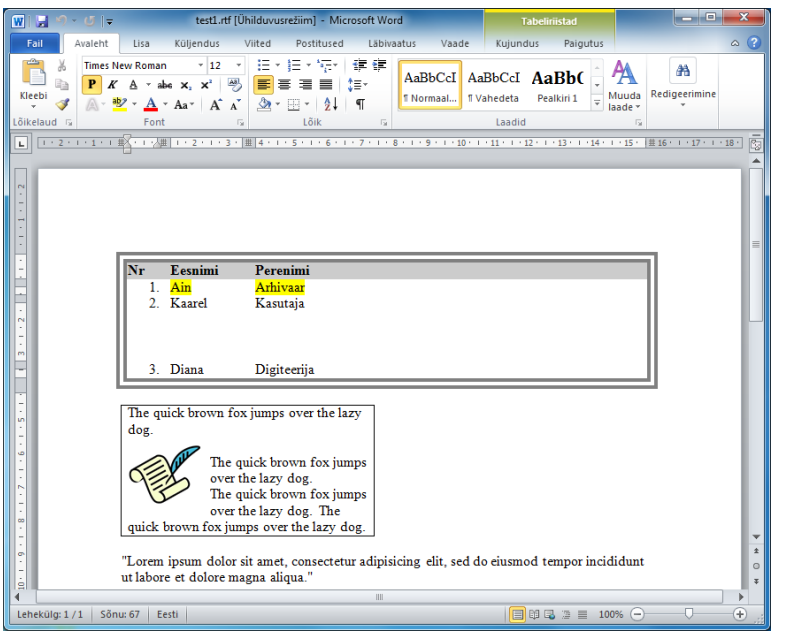
Joonis 4. Hex Editor Neo versioon 5.10 vaade

Avades dokumendi Notepad++ versioon 6.1.1 abil kuvatakse RTFi kodeering, millest on võimalik näha failivormingu ülesehitust (5. joonis).



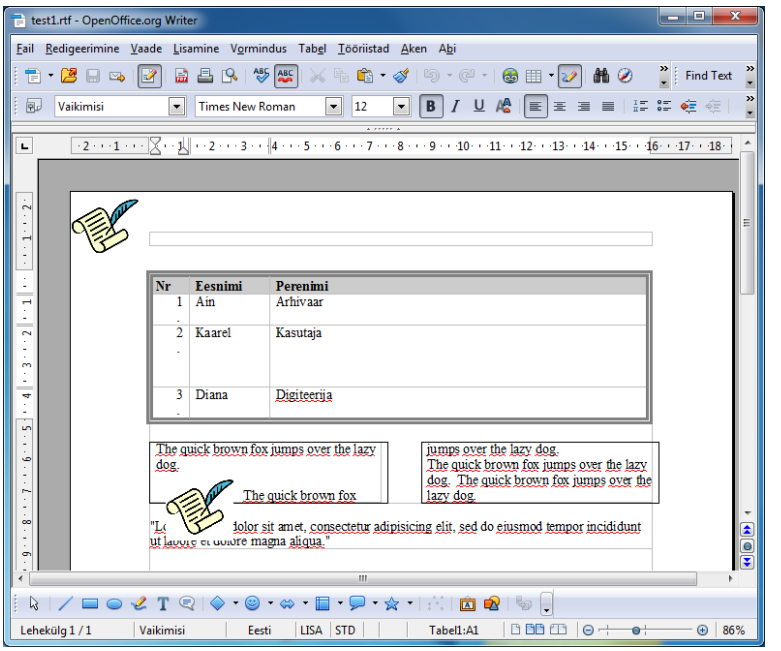
Joonis 5. Notepad++ versioon 6.1.1 vaade

Microsoft Word 2010 avab samuti antud faili (6. joonis), kuid on märgata mõningaid erinevusi (tabelis on puudu ridade raamjooned) võrreldes originaaltarkvara kuvaga.

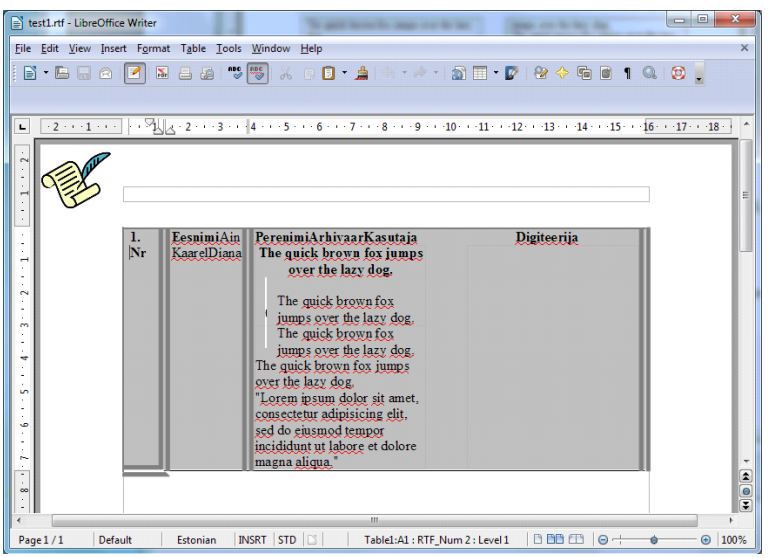


Joonis 6. Microsoft Word 2010 vaade

Kui avada sama fail OpenOffice Writeri (7. joonis) või LibreOffice Writeri (8. joonis) abil, siis on erinevusi veelgi rohkem.



Joonis 7. OpenOffice Writer 3.3.0 vaade



Joonis 8. LibreOffice Writer 3.4.4 vaade

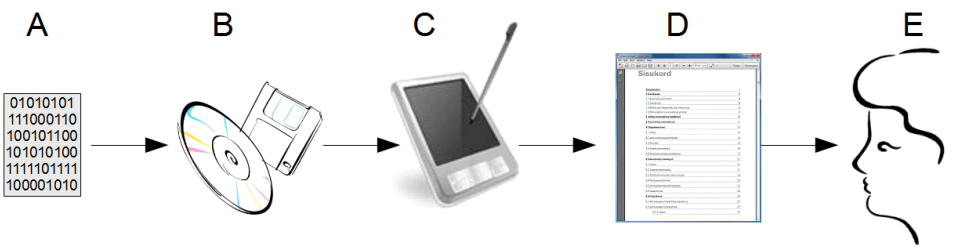
Kõik näitena toodud esitused on tehnoloogilises mõttes õiged. Faili avamisel veateateid ei teki – tarkvarast lähtudes peavadki esitlused sellised olema. Eeltoodud näites on inimsilmale kohe selge, et esitlusega ei ole kõik päris korras. Paljudel juhtudel on aga keeruline või ei ole üldse võimalik avastada, et esituses on midagi valesti. Nt vaatleme tabelit (9. joonis), kus on ühest vormingust teise migreerimise käigus kadunud taustavärv. Esimeses tabelis on summeeritud kõik ühte värvi lahtrid veergude alla. Kui kasutaja näeb ainult viimast (joonisel parempoolset) tulemust, siis esialgset kuju teadmata on äärmiselt keeruline leida summa arvutamisel aluseks olnud liidetavad.

1400	200	300	1100
400	1000	500	1100
800	900	400	100
800	600	700	200
1800	1900	2600	4200

1400	200	300	1100
400	1000	500	1100
800	900	400	100
800	600	700	200
1800	1900	2600	4200

Joonis 9. Taustavärvi kadumine

Eelkirjeldatut kokku võttes saab väita, et digitaalne info on igas mõttes tehnoloogiliselt läbipõimunud – infotehnoloogilisi abivahendeid on vaja nii andmete loomisel, säilitamisel kui ka kasutamisel (10. joonis).



Joonis 10. Näide digitaalse info sõltuvusest

Kõige algelisematel andmetel (A), millel saab olla vaid kaks olekut 0 ja 1, n.ö bittidel, peab olema kindlasti ka füüsiline kandja (B), milleks võib olla näiteks kõvaketas, DVD-plaat, magnetlint vms. Vastav meedium vajab omakorda riistvara, mille külge ühenduda, ja draivereid ühendusprotokollide infoga (C). Mällu laetud andmed serveeritakse uurijale sobiva väljundi ja operatsioonisüsteemi ning rakendusprogrammide abil (D). Kõige viimases etapis toimub informatsiooni tajumine (E), mis sõltub ühelt poolt info autentsest esitamisest, kuid teisalt ka kasutaja teadmiste pagasist.

See, et andmed on säilinud, et neile on võimalik ligi pääseda, lugeda ja esitada, ei garanteeri paraku, et neid õigesti tõlgendatakse ja mõistetakse. Kõige triviaalsem tõlgendamise näide on seotud kuupäevadega. Näiteks kui infosüsteemi on dokumendi registreerimise ajaks märgitud 10/12/09 ja täpsustavat infot ei ole lisatud, siis eestlane võib tõlgendada seda kui 10. detsembrit 2009, ameeriklane kui 12. oktoobrit 2009 ja hiinlane 9. detsembrit 2010. Antud näide

vihjab asjaolule, et peale infotehnoloogiliste aspektide on informatsiooni mõistmisel oluline ka personaalne taju, kontekst ja varasemad teadmised. See tähendab, et erineva taustaga huvigrupid võivad sama informatsiooni erinevalt tõlgendada ja mõista.

Mäluasutuste seas laialdaselt kasutusel olev OAIS-mudel (*Open Archival Information System*) kasutab mõistet lõppkasutajaskond (*Designated Community*), mis on identifitseeritud grupp potentsiaalsetest kasutajatest, kes peaksid olema võimelised konkreetset informatsioonivalimit mõistma (CCSDS 2012, lk 22), kusjuures informatsiooni mõistmine toimub juurdepääsu käigus läbi personaalse tunnetuse ja kogutud teadmuse.

Mäluasutuste suunalt vaadatuna tähendab see huvigruppide selget kaardistamist ja tundmaõppimist. Juurdepääsusüsteemide loomisel on vaja lõppkasutajaskonna teadmust ette näha. Samas on teada, et lõppkasutajaskond ei ole ajas staatiline, mistõttu peavad mäluasutused arvestama ka kirjelduste ajakohastamise vajadusega.

Lisaks on selge, et mida rohkem ajas edasi, seda enam "kolivad" arhivaalid *on-line* süsteemidesse ja internetti, kuid seal ei ole kasutajagruppide piirid enam nii kergelt tajutavad ning gruppide soovidele ja vajadusele on raske vastu tulla. Tänu juurdepääsu lihtsusele võivad mäluasutustes talletatavast informatsioonist huvituma hakata ka need, kes seda varem ei teinud.

Äärmiselt oluliseks muutub ka see, millist teed kasutatakse soovitud tulemuseni liikumiseks – kuidas moodustatakse otsing, kuidas navigeeritakse otsingutulemustes.

Näiteks on kolm isikut: Tartu ülikooli esimese aasta üliõpilane, alles keskkoolis käiv nooruk ja eesti keelt õppiv välistudeng Inglismaalt. Kõik nad suunatakse nt oma tuttava (juhendaja, õpetaja, sõbra) poolt infot koguma arhiivi infosüsteemi. Nende sisemised personaalsed teadmusbaasid on väga erinevad, kuid nad tahavad leida infot sarnasel teemal.

- Esimese aasta üliõpilane (nimetame edaspidi uurijaks) siseneb arhiivi infosüsteemi ja tahab leida infot ühe Rootsi kuninga kohta. Uurija mäletab, et see isik rajas 1632. aastal Tartu ülikooli, kuid ei suuda meenutada, mis oli selle kuninga täpne nimi. Selle asemel, et sisestada otsingusse Gustav II Adolf, avab uurija hoopis otsingu sõnega "Tartu ülikool" ja valib ajaliseks piirajaks aasta 1632 (sisestab senised seonduvad teadmised). Ilmub otsingutulemuste nimekiri, millest ükskõik millise rea peal klikkides näeb ta kirjeldusest, et ülikooli rajaja on Gustav II Adolf.
- Teine kasutaja (keskkooliõpilane) soovib samuti sama kuninga kohta (nt oma referaadi jaoks) infot leida, kuid tema teadmuses ei eksisteeri fakti, et seesama kuningas on kuidagi Tartu ülikooliga seotud. Ta avab isikute otsingulahtri ja trükib sinna teadaoleva nime. Vastuseks on jällegi nimekiri otsitulemustest ja neist ükskõik millise peal klikkides saab ta kirjeldusest lugeda, et seesama Gustav II Adolf on Tartu ülikooli rajaja.
- Kolmas inimene ei oma kahjuks eelnevaid teadmisi. Tema sooviks on Tartu ülikooli ajaloo kohta lihtsalt rohkem teada saada. Kuna ta on teisest kultuurikontekstist, siis valib esmalt arhiivi infosüsteemi keeleks inglise keele ja

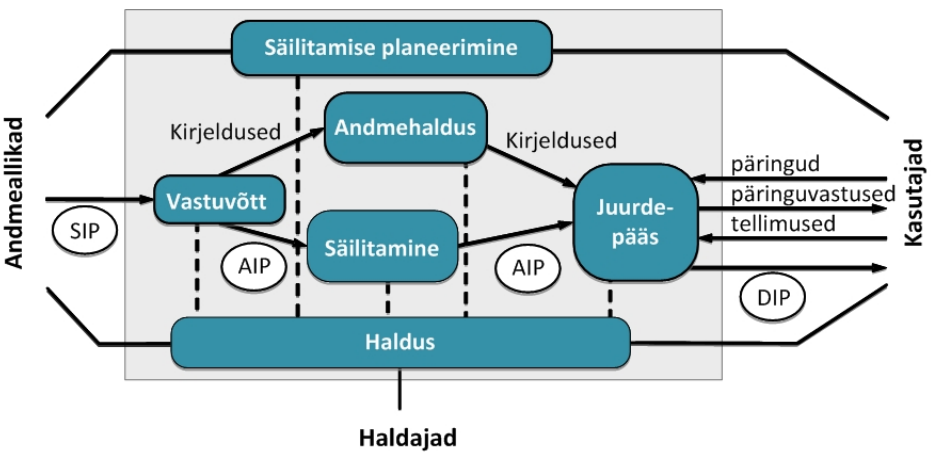
seejärel avab asutuste loetelust Tartu Ülikooli kui asutuse kirje. Talle avaneb vaade, kus on kirjas ka ülikooli kui arhiivimoodustaja ajalugu, millest leiab hõlpsasti, et Tartu ülikooli rajas 1632. aastal Rootsi kuningas Gustav II Adolf.

Hoolimata erinevatest teadmistest jõudsid isikud välja sarnase tulemuseni. Mõni omandas rohkem uusi teadmisi kui teine, kuid informatsiooni tajumise (teadmuse uuenemise) hetkeks olid nad mõistnud esitatavat informatsiooni väga sarnaselt. Sellest lihtsakoelisest näitest peegeldub asjaolu, et laialdase teadmusbaasi ja mitmetahuliste otsinguvõimaluste abil võib olla võimalik mõjutada erinevate inimeste tajusid objektiivse teadmuse tekkimise suunas.

On selge, et lõppkasutajaskonna uurimishuvide rahuldamine eeldab mitmekülgsemaid teadmusbaase, kuid ei ole üheselt fikseeritud, millised peaksid olema teoreetilised alused ja praktilised lahendused nende määratlemiseks.

Seega võib tõdeda, et digitaalarhiivinduse suurim väljakutse on bitijadadele inimliku tähenduse andmine arvestades digitaalse informatsiooni mitmekülgseid interpreteerimisvõimalusi ja kasutaja personaalseid teadmisi ning säilitada seejuures kogutud teabe autentsus.

Seetõttu uuriski käesolev väitekiri võimalusi, kuidas toetada teadmuse (mõistetava informatsiooni) digitaalset arhiveerimist tuginedes eelkõige arhiivinduses laialt levinud OAIS-mudeli funktsionaalsuste (11. joonis) täiendamisele, praktilistele eksperimentidele Rahvusarhiivis ja interdistsiplinaarsetele (nt infotehnoloogias levinud veebianalüütika kombineerimine arhiivinduse põhikontseptsioonidega) võtetele.



Joonis 11. OAIS-mudeli funktsionaalsed kogumid (CCSDS 2012, p 44)[6]

[6] Käesoleva töö autori eestikeelne ümberjoonistus OAIS-mudelist.

Uurimistöö läbiviimiseks kaardistati esmalt uurimisvaldkond. Selleks analüüsiti mäluasutuste seas tunnustatud OAIS-mudelit ja kirjeldati täpsemalt neid funktsionaalseid kogumeid, mis on antud uurimisprobleemi vaatenurgast kõige olulisemad.

Lisaks analüüsiti artikleid, mis käsitlevad digitaalset teadmust ja selle arhiveerimist (ptk 1.1). Kõige rohkem aktuaalset informatsiooni pakkus erinevate digitaalarhiivinduslike projektide tulemuste (dokumentide) läbitöötamine. Nimelt, kuna infotehnoloogia areng mõjutab oluliselt ka digitaalarhiivinduse arengu kiirust, siis digitaalse teadmuse arenduste ja kontseptsioonide viimane terviklik seis kajastub kõige paremini konkreetsete digitaalarhiivinduslike projektide uurimistulemustes mitte niivõrd üksikartiklites või raamatutes. Seetõttu valiti antud uurimistöö jaoks välja 18 digitaalarhiivinduslikku projekti, mille dokumentatsiooni uuriti põhjalikumalt. Projektid valiti välja kindlate tunnuste alusel:

- partnerite arv (valimisse kaasati ainult need rahvusvahelised projektid, kus osales rohkem kui viis partnerit);
- seotus teemaga (ka esmapilgul ainult osaliselt teemaga haakuvad projektid kaasati valimisse, et välistada olulise teabe kõrvalejäämist);
- ajaraam (valiti ainult need projektid, mis lõppesid alles hiljuti või on veel käimas, et vältida vananenud info sattumist valimisse);
- levik, tulemid (valiti ainult need projektid, mille kohta oli loodud avaliku juurdepääsuga dokumente);
- jätkusuutlikkus (valim jäeti lahtiseks, et sinna saaks lisada ka projekte, mis algavad hiljem kui käesolev uurimistöö).

Üldine tegevusskeem jaotati järgnevateks etappideks:

- Selgitada välja digitaalse teadmuse arhiveerimise töövoog, mis arvestaks ka Eestis kehtivaid arhiveerimise nõudeid.

  Töövoog peab olema piisavalt paindlik ja universaalne, et katta isegi hübriidse (nii digitaalsed kui ka paberdokumendid) teabe arhiveerimist. Vastavat töövoogu kirjeldati I artiklis.
- Eestis kehtiva sisendinfopaketi (SIP) spetsifikatsiooni uuendamine, et toetada digitaalse teadmuse arhiveerimist.

  Kavandatud spetsifikatsioon peab olema nii kooskõlas kehtivate nõuetega, kui ka paindlik, et võimaldada spetsiifiliste kirjelduste lisamist, mis aitaksid informatsiooni rikastada. Spetsifikatsiooni tutvustati esimeses artiklis ja ka Rahvusarhiivi vastaval veebilehel (http://www.arhiiv.ee/et/uam).
- Luua tarkvaraline vahend, mis toetab eelnevalt mainitud töövoogu ja SIPi spetsifikatsiooni.

  Selleks, et teoreetilisi tulemeid praktikas katsetada, uuendati Rahvusarhiivi tööriista – universaalset arhiveerimismoodulit (UAM).
- Viia läbi pilootprojekt nii UAMi kui ka spetsifikatsioonide testimiseks.

Pilootprojekt viidi läbi rahvastikuministri büroo materjalide arhiveerimise näol. Projekti käiku kirjeldati esimeses artiklis.

- Selgitada välja vastav Euroopa-ülene parim praktika.
  Detailse ülevaate omamine võimaldab teha muudatusi ja täiendusi varasemas töös. Seetõttu osales käesoleva töö autor ka E-ARKi projekti vastava dokumendi (D3.1 E-ARK Report on Available Best Practices) väljatöötamise juures.
- Töötada välja uus lähenemine lõppkasutajaskonna identifitseerimiseks.
  Kuigi OAIS-mudel toob välja lõppkasutajaskonna olulisuse, ei ole selle määramiseks loodud seni ühtegi head meetodit. Seetõttu analüüsiti teemat põhjalikumalt ja pakuti välja uus meetod, mis põhineb antud töö uuringutel. Vastavat meetodit kirjeldati teises artiklis.
- Luua kasutajakesksed juurdepääsufunktsionaalsused arhiivi infosüsteemi jaoks.
  See uuringu samm keskendus lõppkasutajaskonna teadmusbaasile ja uudsetele juurdepääsulahendustele. Vastavat lahendust kirjeldati kolmandas artiklis.
- Potentsiaalse lahenduse väljatöötamine, mis toetaks arhiivikasutajate teadmiste teisendamist arhiivi teadmuseks.
  Kasutajatel võib olla olulisi teadmisi, mis aitaksid täita lünki juba arhiveeritud infos. Seega uuriti võimalusi sellise teabe arhiivi ülekandmiseks. Vastavaid võimalusi kirjeldati neljandas artiklis.
- Panustada ühtse (rahvusvahelise) sisendinfopaketi spetsifikatsiooni loomisse.
  Mäluasutused kasutavad erinevaid SIPi vorminguid, kuid nende ühtlustamine aitaks saavutada paremat koosvõimet. Selle asemel, et luua täiesti uus spetsifikatsioon või nõuda mõne konkreetse standardi kasutamist, töötas E-ARK välja ühtlustatud versiooni SIPi vormingust, mille kasutuselevõtt ei sunni mäluasutusi oma vorminguid täielikult välja vahetama. Käesoleva töö autor osales spetsifikatsiooni väljatöötamises.

Töö uuris OAIS-mudelit praktilisest vaatenurgast ja laiendas seal esitatud teadmuse käsitlust. Käesolev väitekiri tõi välja OAIS-mudeli vastavad kitsaskohad ja lähtudes OAIS-mudeli funktsionaalsetest kogumitest pakkus välja mõned konkreetsed tehnilised lahendused. Antud dissertatsioon tugines järgmistele uurimisküsimustele:

1. Kuidas on võimalik OAISi vastuvõtufunktsionaalsust (detailsemate kirjelduste ainese ülesehituse, konteksti ja seoste kohta lisamisega) laiendada, et saaks teavikuid juba enne arhiveerimist säilitamiseks ette valmistada ja seeläbi ühtlasi ka informatsiooni digitaalseks teadmuseks muutumisele kaasa aidata?

Uurimus kinnitas vastuvõtu-eelse etapi olulisust – läbiviidud praktiline eksperiment tõestas OAISi vastuvõtu (*Ingest*) funktsionaalsuse laiendamise

võimalikkust ja vastuvõtu-eelse (*Pre-Ingest*) komponendi vajalikkust. Pilootprojektis demonstreeriti ja katsetati erinevaid infotehnoloogilisi võtteid arhiveeritava teabe kvaliteedi tõstmiseks.

Praktiline tulem oli UAMi täiendamine, et see tööriist toetaks väljatöötatud uusi spetsifikatsioone (SIPi vorming ja üleandmise töövoog). Tarkvaralise tööriista oluline positiivne lisaefekt oli selles, et selle abil oli võimalik mitmeid ressursimahukaid (nt kirjelduste valideerimine) tegevusi automatiseerida.

Küsimust käsitleti artiklis "Digital Preservation of Knowledge in the Public Sector: A Pre-Ingest Tool".

Äramärkimist väärib ka asjaolu, et UAM on käesoleval hetkel ainuke tarkvara, mida Rahvusarhiiv soovitab asutustel arhivaalide ettevalmistamiseks ja üleandmiseks kasutada.

2. Millist meetodit saab kasutada arhiivi kasutajaskonna profileerimise toetamiseks?

Arhiveeritud digitaalne teadmus peab olema arhiivi (lõpp)kasutajaskonnale iseseisvalt mõistetav. Seetõttu on oluline teada, kes on arhiivi peamised kasutajad. Kasutajate profileerimiseks saab rakendada doktoritöö raames välja töötatud TRIANGULAR-meetodit. Meetod toetab lõppkasutajaskonna teadmuse kohta rohkema teabe kogumist ja kasutajate teadmusbaasi muutumise jälgimist ajas. Viimane on oluline, sest kasutajate teadmusbaasid võivad muutuda ja arhiveeritud teave ei pruugi seetõttu neile enam mõistetav olla.

Selle uurimisküsimuse käsitlemise praktiline tulem oli näidiskonfiguratsioon, mis põhines vabavaralisel analüütikaplatvormil "Piwik" (https://piwik.org/). Platvorm on unikaalne, sest on vabavaraline ja võimaldab mäluasutustel endil oma kogutud informatsiooni täielikult kontrollida, s.t et mäluasutused ei pea kogutud infot kolmanda osapoole juures hoidma, seega ei riiva antud platvormi kasutamisega oma kasutajate privaatsust.

TRIANGULAR-meetod on väga paindlik ja võimaldab koguda infot üsna automaatselt, mis tähendab, et tegemist on kuluefektiivse lahendusega – peamine ressurss kulub lahenduse käivitamiseks ja tulemuste töötlemiseks.

OAIS-mudeli seisukohast vaadatuna toetab TRIANGULAR-meetod *Monitor Designated Community* ja *Monitor Technology of the Preservation Planning* funktsionaalsusi, pakkudes praktilise lahenduse lõppkasutajaskonna identifitseerimiseks ja haldamiseks.

Küsimust käsitleti artiklis "Digital Preservation and Knowledge in the Public Archives: For Whom".

3. Kuidas saab mitmetahuline juurdepääs toetada arhiveeritud teadmuse kasutamist?

Juurdepääsusüsteemide kasutusliidesed peavad olema väga paindlikud ja mugavad ning võtma sealjuures arvesse ka kasutajaskonna teadmusbaasi. Klassikaline tekstipõhine otsing töötab hästi ainult homogeense teadmusbaasiga

kasutajate korral. Kui kasutajad teavad "õigeid" sõnu, siis jõuavad nad lõpuks soovitud tulemuseni. Samas, kuna kasutajatel on erinevad kogemused, teadmised ja huvid, siis peab ka otsingu alustamiseks olema rohkem kui üks võimalus. Lisaks klassikalisele otsingule peaks olema ka tsüklilise otsimise võimalus – otsingut alustatakse laia valimiga, mis ei nõua kasutajalt väga põhjalikke teadmisi, ja siis kitsendatakse otsitulemuste arvu jooksvalt täpsustavate filtrite rakendamisega. Selline juhendav otsimise funktsionaalsus võimaldab navigeerida tulemuseni ka ilma ühtegi otsisõna sisestamata.

Tavaliselt on arhiivides loodud ka sisuspetsiifilised juurdepääsusüsteemid. Nt fotode otsimiseks ja kuvamiseks on üks süsteem, filmide otsimiseks ja vaatamiseks teine. Ühelt poolt võimaldab selline lähenemine lihtsamini uusi süsteeme luua, aga teisalt pakub see ka kasutajatele mugavamat sisuga tutvumise võimalust (nt fotode ja filmide jaoks on erinevad, tüübispetsiifilised vaaturid). Samas on selle arhitektuuri juures ka üks suur probleem – reeglina tuleb otsingud teha kõikides süsteemides eraldi. Lahenduseks on objektide omavaheline linkimine, mis võimaldab arhiivi kesksest infosüsteemist teostada otsinguid ka sisuspetsiifiliste süsteemide kohta ja soovitud tulemuse leidmisel liikuda edasi juba vastavasse eraldiseisvasse süsteemi.

Samas, taolise mitmetahulise otsingu- ja sirvimisfunktsionaalse lisamine juurdepääsusüsteemile võib olla mäluasustuste jaoks suur väljakutse, sest see võib nõuda palju eeltööd (andmete korrastamist, taksonoomiate moodustamist, andmete rikastamist jms). Nii oli see ka Rahvusarhiivis, kus ühiste jõupingutuste abil valmisid klassifikaatorite taksonoomiad, mida oli võimalik seejärel integreerida Rahvusarhiivi juurdepääsu infosüsteemi uude versiooni.

Uurimisküsimuse raames viidi läbi mitmetahuliste otsi- ja sirvimisvõimaluste arendamine ning rakendamine Rahvusarhiivi juurdepääsu infosüsteemis – praktiline tulem oli arhiiviportaali uuendus, mis toetab uusi võimalusi.

Küsimust käsitleti artiklis "Toward Common Ontologies of Facets of the Archival Access Portal". OAIS-mudeli mõistes pakub see käsitlus tuge *Coordinate Access Activities* funktsioonide rakendamisele praktikas.

4. Kuidas saavad kasutajad oma teadmisi arhiiviga jagada?

On tõenäoline, et arhiivimoodustajad teavad oma informatsiooni kohta rohkem kui teised ja seetõttu saaksid nad ka potentsiaalselt kõige efektiivsemalt arhiveerimisele kuuluvat informatsiooni rikastada. Seda saab teha näiteks UAMi abil, kui informatsiooni arhiveerimiseks ette valmistatakse, nii jõuaks informatsioon arhiivi juba rikastatud kujul.

Samas on arhiivides (k.a Rahvusarhiivis) palju sellist informatsiooni, mida oleks samuti mõistlik lubada huvitatud osapooltel rikastada. Esialgsete katsetuste läbiviimiseks otsustati defineerida kõik rikastamiseks väljajagatavad objektid omaette ressurssidena ja publitseerida masintöödeldavate avaandmetena. Taoline lähenemine võimaldab mitte ainult ühispanustamist, vaid ka uute tehnoloogiate rakendamist (nt mustripõhine andmekaeve), mis varasemalt olid

kasutusel vaid teistes distsipliinides. Oluline on märkida, et lisaks arhiivile saavad kõik teisedki uusi rakendusi avatud teabe peale üles ehitada.

Selleks töötati välja esialgne vahendite kogum, mis võimaldab arhiiviportaalist andmete suunamist ühispanustamise rakendustesse ja rikastatud andmete saatmist tagasi portaali. Hoolimata sellest, et antud töö käigus saavutatud tulemus on suures osas veel *proof-of-concept*-staadiumis, on tegemist siiski suure edasiminekuga. Lisaks võib olulise punktina tuua välja asjaolu, et antud väljakutse lahendamise raames jõuti ettepanekuni kasutada arhiivis paikneva teabe rikastamiseks OAIS-mudeli *Ingest* komponendi asemel *Receive Database Update* funktsiooni.

Seda teemat käsitleti artiklis "Transforming users' knowledge into archival knowledge".

Uurimisteemad edaspidiseks

On täiesti selge, et teadmuse digitaalne arhiveerimine ei ole pelgalt seisund või koht, kuhu mingil hetkel jõutakse. See on kulgev teekond, kombinatsioon pidevast seirest ja muutuste haldamisest, mistõttu ei taotle käesolev dissertatsioon kõikide teadmuse digitaalse arhiveerimisega seotud aspektide lahendamist, vaid käsitleb, nagu eelnevast näha, ainult selle valdkonna valitud kitsaskohti.

Käesolev töö leidis vastused mitmele olulisele digitaalse arhiveerimisega seotud küsimusele, kuid mitmed potentsiaalselt lahendamist vajavad ülesanded on veel ees:

- Uurida SIPi vormingu rakendamist ka teistes valdkondades. Näiteks äri- ja personaalarhiveerimise valdkonnas vajatakse samuti digitaalse teadmuse arhiveerimise vormingut.
- Uurida kasutajate profileerimist praktikas. See ülesanne nõuab prototüüpide loomist ja nende testimist reaalses keskkonnas.
- Mitmetahulisi otsinguid luues tuleks laiendada seoseid teiste olemasolevate taksonoomiatega. See aitaks ära hoida olukorra, kus arhiveeritud informatsioonist saab ühel hetkel "üksik saar". Teadmus on oma olemuselt dünaamiline ja võiks omada seoseid ka teiste arhiivikogudega.
- Uurida täpsemalt ühispanustamisest tekkivat kasu. Kui ühispanustamise kasulikkuse kohta on võimalik tuua välja konkreetsed numbrid, siis võib see julgustada mäluasutusi oma kogusid rohkem jagama.
- Arendada välja konkreetne standard arhiivi kataloogi ja ühispanustamise rakenduste vaheliseks andmevahetuseks.

REFERENCES

Advanced Workflow Preservation Technologies for Enhanced Science (Wf4Ever). Accessed 29 November 2015 at: http://www.wf4ever-project.org/

Alliance Permanent Access to the Records of Science in Europe (APARSEN). Accessed 29 November 2015 at: http://www.alliancepermanentaccess.org/

Andrews, Jean. *A+ Guide to Managing and Maintaining Your PC*, 2009, pp 135.

ARchive COmunities MEMories (ARCOMEM). Accessed 29 November 2015 at: http://www.arcomem.eu

Archives Portal Europe (APE), Directory, 2016. Accessed 29 November 2015 at: www.archivesportaleurope.net/directory

BlogForever. Accessed 29 November 2015 at: http://blogforever.eu/

Borghoff, Uwe. M; Rödig, Peter; Scheffczyk, Jan; Schmitz, Lothar. *Long-Term Preservation of Digital Documents*. Springer-Verlag Berlin Heidelberg, 2006.

Collaboration to Clarify the Costs of Curation (4C). Accessed 29 November 2015 at: http://4cproject.eu/

Consultative Committee for Space Data Systems. *Reference model for an Open Archival Information System (OAIS) (Magneta Book CCSDS 650.0-m-2)*, 2012. Accessed 08 April 2016 at: http://public.ccsds.org/publications/archive/650x0m2.pdf

Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval (CASPAR). Accessed 29 November 2015 at: http://www.casparpreserves.eu/

Dobreva, Milena. *On the Knowledge Management Aspects of Digital Preservation*, 2009. Accessed 08 April 2016 at: https://research.uni-sofia.bg/bitstream/10506/645/1/S3T2009_28_Milena_Dobreva.pdf

Electronic Knowledge Management Project Team (EKMPT), 2010, Electronic Knowledge Management (EKM), A Framework for Understanding EKM with Guidelines for Retaining Intellectual Capital and Addressing eDiscovery Requirements. Accessed 08 April 2016 at: http://www.aia-aerospace.org/assets/report_ekm_052010.pdf

EMC Digital Universe, IDC. The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things, 2014. Accessed 08 April 2016 at: http://www.emc.com/leadership/digital-universe/2014iview/it-imperatives.htm

Enabling kNowledge Sustainability Usability and Recovery for Economic (ENSURE). Accessed 29 November 2015 at: http://www.prelida.eu/ensure-enabling-knowledge-sustainability-usability-and-recovery-economic-value

European Archival Records and Knowledge Preservation (E-ARK). Accessed 08 April 2016 at: http://www.eark-project.com/

ForgetIT. Concise Preservation by combining Managed Forgetting and Contextualized Remembering. Accessed 29 November 2015 at: http://www.forgetit-project.eu/en/

Giaretta, David. *Advanced Digital Preservation*. Springer, 2011.

Giaretta, David. D2101: Prototype OAIS-infrastructure, 2008. Accessed 01 September 2014 at: http://casparpreserves.eu/Members/cclrc/Deliverables/prototype-oais-infrastructure/at_download/file.pdf

Giaretta, David. E-Infrastructures and long-term preservation. Scientific Publishing in the European Research Area, 2007. Accessed 08 April 2016 at: ec.europa.eu/research/science-society/document_library/pdf_06/giaretta-022007_en.pdf

Interpares, Emulation, 2016. Accessed 29 November 2015 at: http://www.interpares.org/ip2/ip2_term_fdisplay.cfm?tid=550&pageID=1

IT Business Edge. Static Electricity and Computers, 2003. Accessed 08 April 2016 at: http://www.webopedia.com/DidYouKnow/Computer_Science/static.asp

Johnston, Lisa. Open Archival Information System (OAIS). An Introduction, 2013. Accessed 08 April 2016 at: https://www.statearchivists.org/download_file/view/355/635/

Kärberg, Tarvo; Bredenberg, Karin; Skog, Björn; Nielsen, Anders Bo; Johansen, Kathrine Hougaard Edsen; Silva, Hélder. D3.2 E-ARK SIP Draft Specification, 2015. Accessed 08 April 2016 at: http://eark-project.com/resources/project-deliverables/17-d32-e-ark-sip-draft-specification

Kärberg, Tarvo; Bredenberg, Karin; Skog, Björn; Nielsen, Anders Bo; Johansen, Kathrine Hougaard Edsen; Silva, Hélder; Zavrsnik, Gregor; Szilágyi, Levente; Tømmerholt, Phillip Mike. D3.3 E-ARK SIP Pilot Specification, 2016. Accessed 08 April 2016 at: http://eark-project.com/resources/project-deliverables

Kärberg, Tarvo; Oolu, Karin; Randmäe, Piret; Johansen, Kathrine Hougaard Edsen; Thirifays, Alex; Domajnko, Boris; Delve, Janet; Anderson, David. D3.1 E-ARK Report on Available Best Practices, 2014. Accessed 08 April 2016 at: http://eark-project.com/resources/project-deliverables/6-d31-e-ark-report-on-available-best-practices

Keeping Emulation Environments Portable (KEEP). Accessed 29 November 2015 at: http://www.keep-project.eu/ezpub2/index.php

Living Web Archives (LIWA). Accessed 29 November 2015 at: http://liwa-project.eu/

Moore, Reagan. W. *The Preservation of Data, Information, and Knowledge*. San Diego Supercomputer Center, 2001. Accessed 07 July 2010 at: www.sdsc.edu/NARA/Publications/Web/moore-rw.doc

Mueller, Scott. *Upgrading and repairing laptops*, Que Publishing, 2004, pp 422.

NAE. Potential of Agents in Digital Preservation, 2010. Accessed 08 April 2016 at: http://www.arhiiv.ee/files/Potential%20of%20agents%20in%20DP.pdf

Nonaka, Ikujiro. *A Dynamic Theory of Organizational Knowledge Creation*. Organization Science, 5(1): 14-37, February, 1994. Accessed 08 April 2016 at: http://business.illinois.edu/josephm/BA504_Fall%202008/Uploaded%20in%20Nov%202007/Nonaka%20%281994%29.pdf

Preservation and Long-term Access through Networked Services (PLANETS). Accessed 29 November 2015 at: http://www.planets-project.eu/

PReservation Organizations using Tools in AGent Environments (PROTAGE). Accessed 29 November 2015 at: http://www.ra.ee/protage

PREserving Linked DAta (PRELIDA). Accessed 29 November 2015 at: http://www.prelida.eu/

PrestoPRIME. Accessed 29 November 2015 at: http://www.prestoprime.org/

Rahvusarhiiv, Digitaalarhiivi infosüsteem, arhitektuuri kirjeldus, 2008a. Accessed 08 April 2016 at: http://www.ra.ee/public/Digiarhiiv/da_arhitektuur.pdf

Rahvusarhiiv, Digitaaldokumentide arhiveerimise nõuded, 2008b, lk 16. Accessed 29 November 2015 at: http://www.arhiiv.ee/public/Juhised/digidok_arhiveerimine.pdf

Rothenberg, Jeff. Ensuring the Longevity of Digital Information, 1999, pp 4. Accessed 08 April 2016 at: http://www.clir.org/pubs/archives/ensuring.pdf

Scalable Preservation Environments (SCAPE). Accessed 29 November 2015 at: www.scape-project.eu/

Stuchell, Lance T. What is Digital Preservation? 2013. Accessed 08 April 2016 at: http://www.lib.umich.edu/preservation-and-conservation/digital-preservation/what-digital-preservation

Sustaining Heritage Access through Multivalent Archiving (SHAMAN). Accessed 29 November 2015 at: http://www.shaman-ip.eu/

Tessella, British Library. Survey Analysis Report, 2009. Accessed 08 April 2016 at: http://www.planets-project.eu/docs/reports/planets-survey-analysis-report-dt11-d1.pdf

The National Digital Stewardship Alliance (NSDA), NDSA Glossary, 2015. Accessed 29 November 2015 at: www.digitalpreservation.gov/ndsa/ndsa-glossary.html

Thomaz, Katia P. *Critical Factors for Digital Records Preservation*, Journal of Information, Information Technology, and Organizations, Volume 1, 2006

Timeless Business Processes and Services (TIMBUS). Accessed 29 November 2015 at: http://timbusproject.net/

Tzitzika, Yannis; Christophides, Vassilis; Kotzino, Dimitris; Antonio, Grigoris. D2101B: Associated draft reports of knowledge management architecture and tools, 2008. Accessed 01 September 2014 at: http://casparpreserves.eu/Members/cclrc/Deliverables/associated-draft-reports-of-knowledge-management-architecture/at_download/file.pdf

Tzitzika, Yannis; Christophides, Vassilis; Kotzino, Dimitris; Marketakis, Yannis; Antonio, Grigoris. D2102: Prototype of registry-related KM services, 2008. Accessed 01 September 2014 at: http://casparpreserves.eu/Members/cclrc/Deliverables/prototype-of-registry-related-km-services-1/at_download/file.pdf

Vallaste, Heikki. Emuleerimine, Vallaste e-teatmik, 2012. Accessed 29 November 2015 at: www.vallaste.ee

ACKNOWLEDGEMENTS

First of all, I would like to thank my supervisor, Professor Aadu Must, for his professional guidance and support during the doctoral studies.

I would also like to thank the Digital Archives at the National Archives of Estonia – especially Director of the Digital Archives Mr Lauri Leht for his continuous support – my colleagues at the National Archives of Estonia for their enthusiastic discussions, and those involved in the E-ARK project for their eye-opening ideas.

I am also very grateful to State Archivist Mr Priit Pirsko for his initiative to start my research in the digital preservation area and to a good friend, Mr Koit Saarevet, who has successfully managed several important digital preservation projects. I would also like to credit a colleague from the National Archives of Estonia, Helina Tamman for editing the summary in Estonian of this thesis.

I would like to sincerely thank Professor Janet Delve and Professor David Anderson from the University of Brighton for their guidance and support.

I am also very grateful to Professor Kurmo Konsa from the Tartu Art College (University of Applied Sciences) for his mentoring and to Aigi Rahi-Tamm, Associate Professor in Archival Studies at the Institute of History and Archaeology of the University of Tartu for her great coordination and management.

Special gratitude goes to my dear wife Kati. She has always been so inspiring and kind in listening to all of my new ideas and worries. I gratefully acknowledge her continuous support and admire her purposefulness.

Many thanks to other good people for their positive attitude, as well as for their constructive criticism, without which my research journey would not have been possible.

PUBLICATIONS

APPENDICES

Appendix A. The full version of project cards

3.1.1.1. 4C

Name	Collaboration to Clarify the Costs of Curation (4C)
Web page	http://4cproject.eu/
Funded by	European Union under the Seventh Framework (FP7) programme for research and technological development and demonstration activities
Duration	From 2013 to 2015
Partners	• Jisc (UK) (Project Co-ordinator) • The Royal Library – National Library of Denmark (Denmark) • INESC-ID – Institute for System and Computer Engineering (Portugal) • Danish National Archives (Denmark) • Deutsche Nationalbibliothek (Germany) • Humanities Advanced Technology and Information Institute (HATII) - University of Glasgow (UK) • University of Essex (UK) • Keep Solutions (Portugal) • Digital Preservation Coalition (UK) • Secure Business Austria (Austria) • Digital Curation Centre – University of Edinburgh (UK) • DANS – Data Archiving and Networked Services (KNAW/NWO) (Netherlands) • National Library of Estonia (Estonia)
Description	4C helps organisations across Europe to invest more effectively in digital curation and preservation. Research in digital preservation and curation has tended to emphasize the cost and complexity of the task in hand. 4C reminds us that the point of this investment is to realise a benefit, so our research must encompass related concepts such as ‘risk’, ‘value’, ‘quality’ and ‘sustainability’.
Progress	D1.2 Sustainability & Benefits Realisation Plan D2.1 Baseline Study of Stakeholder & Stakeholder Initiatives D2.2 Stakeholder Registry D2.3 Final Stakeholder Report D2.4 Final Report on Outreach Events D2.5 Project Communication Plan D2.6 Report on Communications Activities D2.8 Curation Costs Exchange D3.1 Evaluation of Cost Models and Needs & Gaps Analysis D3.1 Summary of Cost Models D3.2 Cost Concept Model and Gateway Specification D3.3 Curation Costs Exchange Framework D4.1 A Prioritised Assessment of The Indirect Economic Determinants of Digital Curation

	D4.2 Assessment of Community Validation of the Esrm D4.3 Quality and Trustworthiness as Economic Determinants in Digital Curation D4.4 Report on Risk, Benefit, Impact and Value D4.5 From Costs to Business Models D5.1 Draft Roadmap D5.2 Roadmap Report
Relevance	The project was focused on estimating and comparing the cost of digital curation (as well digital preservation). Although the cost component is gathering more and more importance in the digital preservation area, the 4C project will not be further investigated as the curation cost modelling is out of the scope of the research planned for this dissertation.
Note	All information represented here (except Relevance) has been harvested from the web page of this project.

3.1.1.2. APARSEN

Name	Alliance Permanent Access to the Records of Science in Europe (APARSEN)
Web page	http://www.alliancepermanentaccess.org/
Funded by	European Commission, Seventh Framework (FP7) ICT Programme
Duration	From 2011 to 2014
Partners	Coordinator • Science and Technology Facilities Council (United Kingdom) Participants • Verein Zur Forderung Der It-Sicherheit in Osterreich (Austria) • Oesterreichische Nationalbibliothek (Austria) • European Organization for Nuclear Research (Switzerland) • Alfred-Wegener-Institut Helmholtz- Zentrum Fuer Polar- und Meeresforschung (Germany) • Deutsche Nationalbibliothek (Germany) • Globit-Globale Informationstechnik Gmbh (Germany) • Ftk Forschungsinstitut fur Telekommunikation und Kooperation Ev (Germany) • Incontec Gmbh (Germany) • Inmark Estudios Y Estrategias Sa (Spain) • Csc-Tieteen Tietotekniikan Keskus Oy (Finland) • Airbus Operations Sas (France) • European Space Agency (France) • Cines (France) • University of Patras (Gr) • Foundation for Research and Technology Hellas (Gr) • Ibm Israel - Science and Technology Ltd (Israel) • Fondazione Rinascimento Digitale-Nuove Tecnologie Per I Beni Culturali (Italy) • Universita Degli Studi Di Trento (Italy) • Consorzio Interuniversitario Nazionale Per L'informatica (Italy)

	• Koninklijke Nederlandse Akademie Van Wetenschappen – Knaw (Netherlands) • Stichting Secretariaat Van De International Association of Scientific, Technical & Medical Publishers (Netherlands) • Stichting European Alliance for Permanent Access (Netherlands) • Stichting Liber (Netherlands) • Koninklijke Bibliotheek (Netherlands) • Lulea Tekniska Universitet (Sweden) • University of Essex (United Kingdom) • Tessella Ltd (United Kingdom) • The British Library Board (United Kingdom) • Digital Preservation Coalition Limited By Guarantee*Dpc (United Kingdom)
Description	APARSEN was a Network of Excellence that brought together an extremely diverse set of practitioner organisations and researchers in order to bring coherence, cohesion and continuity to research into barriers to the long-term accessibility and usability of digital information and data, exploiting diversity by facilitate the building of a long-lived Virtual Centre of Digital Preservation Excellence. The objective of this project may be simply stated, namely to look across the excellent work in digital preservation which is carried out in Europe and to try to bring it together under a common vision. With the APARSEN project, the APA has a blueprint for a Centre of Excellence which brings together expertise from the pioneers in digital preservation on the basis of an integrated view of digital preservation techniques. The term "Virtual Centre of Excellence" is used to emphasise the point that this does not depend on the expertise of a single organisation. Solutions to problems in digital preservation are addressed through consultancy, training and tools and services.
Progress	D11.6 Virtual Centre of Excellence development D13.3 Report about Standardisation Activities D45.2 Mapping defragmentation opportunities across key stakeholders outside APARSEN D42.2 Report on definition of curricula and course contents, summary of liaison actions and co-operations D22.4 The Interoperability Framework Implementation with Added Value Services D11.5 Report on a Common Vision of Digital Preservation D43.2 Report on launch of digital preservation training portal for VCoE D31.1 Report on DRM preservation D22.3 Demonstrator set up and definition of added value services: Part 2 D11.4 Virtual Centre of Excellence Development: Progress to Year 3 D13.2 Report about Standardisation Activities: Progress to Year 3 D11.3 Report on a Common Vision for Digital Preservation: Progress to Year 3 D44.3 Report on interactive map of stakeholders in digital preservation D36.2 Exemplar business cases

	D32.2 Report on testing of cost models and further analysis of cost parameters D27.1 Recommendations about scalability D23.2 Storage solutions analysis report D25.2 Interoperability Strategies D13.1 Report about standardisation activities: progress to year 2 D42.1a Annex for Report on existing initiatives and curricula regarding digital long term preservation
Relevance	The project was focused on identifying and merging research areas of digital preservation into a common vision. APARSEN addresses solutions to problems in digital preservation through consultancy, training, tools and services. As the dissertation deals with more detailed and focused research questions (more specific to digital preservation of knowledge) then the project will not be further investigated.
Note	All information represented here (except Relevance) has been harvested from the web page of this project or from the CORDIS system* (http://cordis.europa.eu/).

3.1.1.3. ARCOMEM

Name	ARchive COmunities MEMories (ARCOMEM)
Web page	http://www.arcomem.eu
Funded by	European Commission, Seventh Framework (FP7) ICT Programme
Duration	From 2011 to 2013
Partners	• University of Sheffield, • Internet Memory Foundation,, • University of Southampton • Athena Research and Innovation Center in ICKT • Institut Télécom, Télécom ParisTech • Deutsche Welle, Südwestrundfunk, • Yahoo! Iberia • L3S Research Center • Hellenic Parliament • Austrian Parliament • Athens Technology Center SA
Description	Aim was to: • help transform archives into collective memories that are more tightly integrated with their community of users • exploit Social Web and the wisdom of crowds to make Web archiving a more selective and meaning-based process Innovative tools for archivists were provided to help exploit the new media and make our organisational memories richer and more relevant. It was done in three ways: • showed how social media can help archivists select material for inclusion, providing content appraisal via the social web • showed how social media mining can enrich archives, moving towards structured preservation around semantic categories

	• looked at social, community and user-based archive creation methods As results of this activity the outcomes of the ARCOMEM project included: • innovative models and tools for Social Web driven content appraisal and selection, and intelligent content acquisition • novel methods for Social Web analysis, Web crawling and mining, event and topic detection and consolidation, and multimedia content mining • reusable components for archive enrichment and contextualization • two complementary example applications, the first for media-related Web archives and the second for political archives a standards-oriented ARCOMEM demonstration system
Progress	D1.1 User requirements and System Architecture D1.2 Updated User requirements and System Architecture D2.1 Scenarios and Social Media Datasets D2.2 Social Web-based Archive Contextualization V1 D2.3 Social Web-based Archive Contextualization V2 – with update D3.1 Models and Architecture Definition D3.2 Extraction and Enrichment D3.3 Extraction and Enrichment – with update D4.2 Opinion Mining V1 D4.3 Opinion Mining v2 – with update D5.1 Models and Architecture Definition (Intelligent Content Acquisition) D5.2 Intelligent and adaptive content acquisition V1 D6.1 Requirements and system architecture D7.1 ARCOMEM Development Guideline D7.2 Integration Prototype V1 D7.3 ARCOMEM API Reference Guide D8.1 ARCOMEM Archivist's Tool V1 Report D8.2 Documentation and Evaluation of the ARCOMEM Archivist's Tool V1 D9.1 ARCOMEM Parliament Tool V1 Report D10.1 Community Demonstrator "Intelligent Content Acquisition"
Relevance	The project was focused on social web, automated information creation and appraisal. Some reusable software components to support the collection, analysing and access of content from Web and Social Web have been delivered.[7] Although the project took the leverage of wisdom of crowds, it still has a different focus compared to this research. Therefore, as the dissertation deals with research questions more specific to digital preservation of knowledge, then the project will not be further investigated.
Note	All information represented here (except Relevance) has been harvested from the web page of this project or from the CORDIS system* (http://cordis.europa.eu/).

[7] ARCOMEM Semantic and social web crawling, http://sourceforge.net/projects/arcomem/

3.1.1.4. BlogForever

Name	BlogForever
Web page	http://blogforever.eu/
Funded by	European Commission, Seventh Framework (FP7) ICT Programme
Duration	From 2011 to 2013
Partners	Coordinator • Aristotelio Panepistimio Thessalonikis (Gr) Participants • European Organization for Nuclear Research (Switzerland) • Technische Universitat Berlin (Germany) • Populis Gmbh (Germany) • Anonymos Etairia Erevnas Anaptyxis Neon Technologion Kai Efarmogon Internet (Gr) • Altec Software Development S.A.(Gr) • Efstratios Arampatzis Monoprosopi Epe (Gr) • Cyberwatcher (Norway) • Srdc Yazilim Arastirma Ve Gelistirme Ve Danismanlik Ticaret Limited Sirketi (Turkey) • University of Glasgow (United Kingdom) • The University of Warwick (United Kingdom) • University of London (United Kingdom)
Description	BlogForever, a collaborative European Commission funded project, developed an exciting new system to harvest, preserve, manage and reuse blog content. The system is performing an intelligent harvesting operation which retrieves and parses hypertext as well as all other associated content (images, linked files, etc) from blogs. It copies content by interrogating not only the RSS feed of a blog, but also by copying data from the original HTML. The parsing action is able to render the captured content into structured data, expressed in XML; it does this in accordance with the project's data model. The result of this action is carving semantic entities out of blog content on an unprecedented micro-level. Author names, comments, subjects, tags, categories, dates, links, and many other elements are expressed within a hierarchical structure. This content is imported into the BlogForever repository (based on CERN's Invenio platform), a public-facing web archiving mechanism which provides facilities to preserve, view, interrogate & reuse the content to an unprecedented degree of detail.
Progress	D2.1 Survey Implementation Report, 2011/08/31, D2.2 Weblog Data Model, 2011/10/31, D2.3 Weblog Ontologies, 2012/05/31, D2.4 Weblog Spider Prototype and Associated Methodology, 2011/11/31 D2.5 Weblog Spam Filtering Report and Associated Methodology, 2012/02/29 D2.6 Data Extraction Methodology, 2012/05/31. D3.1 Preservation Strategy, 2012/09/30, D3.2 Interoperability Prospects, 2013/04/30,

	D3.3 Digital Rights Management Policies, 2013/08/31. D4.1 User requirements and platform specifications, 2011/12/30, D4.2 Weblog spider component design, 2012/06/30, D4.3 Initial weblog spider prototype, 2012/09/30, D4.4 Digital repository component design, 2012/11/30, D4.5 Initial weblog digital repository prototype, 2013/01/31, D4.6 Final weblog spider component, 2013/04/30, D4.7 Final weblog digital repository component, 2013/05/31, D4.8 Final BlogForever platform, 2013/08/30. D5.1 Design of Case Studies, 2012/06/30, D5.2 Implementation of Case Studies, 2013/05/31, D5.3 User Questionnaires and Reports, 2013/05/31, D5.4 System logs, 2013/05/31, D5.5 Case studies comparative analysis and conclusions, 2013/05/31. D6.1 Knowledge Portal , 2011/03/30, D6.2.1 Dissemination Plan (first version). 2011/08/31, D6.2.2 Dissemination Plan (second version), 2012/02/29, D6.3 Market Analysis, 2012/06/30, D6.4 Business Model, 2013/01/31, D6.5 Business Exploitation Plan, 2013/08/31, D6.6 Final Report on Dissemination Activities, 2013/08/31.
Relevance	The project was focused on harvesting, preserving, managing and reusing blog content. More specifically, the system provided by BlogForever retrieves, parses hypertext and associated content (e.g. images) from blogs and is then carving semantic entities (e.g. author names) out of the parsed content. Although the tools and methods provide good means to harvest semantic descriptions about the content, they have still too strong focus on blogs. As this dissertation seeks solutions for more universal content, then the BlogForever project will not be further investigated.
Note	All information represented here (except Relevance) has been harvested from the web page of this project or from the CORDIS system* (http://cordis.europa.eu/).

3.1.1.5. CASPAR

Name	Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval (CASPAR)
Web page	http://www.casparpreserves.eu/
Funded by	European Commission, Sixth Framework (FP6) IST Work Programme
Duration	From 2006 to 2009
Partners	Coordinator • The Science and Technology Facilities Council (United Kingdom) Participants • Ciant – Mezinarodni Centrum Pro Umeni A Nove Technologie V Praze (Czech Republic) • United Nations Educational, Scientific and Cultural Organization – Unesco (France)

	• Centre National De La Recherche Scientifique (France) • Institut National De L'audiovisuel (France) • Institut De Recherche Et De Coordination Acoustique Musique – Ircam (France) • European Space Agency (France) • Foundation for Research and Technology Hellas (Gr) • Ibm Israel - Science and Technology Ltd (Israel) • Advanced Computer Systems A.C.S. S.P.A. (Italy) • Metaware Societa Per Azioni (Italy) • Consiglio Nazionale Delle Ricerche (Italy) • Universita Degli Studi Di Urbino Carlo Bo (Italy) • Asemantics Srl Uninominale (Italy) • Engineering - Ingegneria Informatica Spa (Italy) • University Of Leeds (United Kingdom) • The University Of Glasgow (United Kingdom)
Description	Work in CASPAR was driven by the following objectives: • enhance techniques for capturing representation information and other preservation related information; • design virtualisation services for preserving resources despite changes in computer hardware and software and storage systems; • ensure trustworthiness of preserved data with standard features for digital rights management, authentication, and accreditation; • research more sophisticated access to and use of preserved digital resources such as intuitive query and browsing mechanisms.
Progress	D1101 Review of State of The Art D1201 Caspar Conceptual Model – Phase 1 D1202 Caspar Guidelines D1301 Caspar Overall Component Architecture and Component Model D1302 CASPAR System Development – Overall Master Plan D1203 Report on Trusted Digital Repositories D1303 CASPAR – Overall Architecture, Components and Interfaces D2103 Prototype of Descriptive Information-related KM Services D2101 Prototype OAIS Infrastructure D2102 Prototype of registry-related KM services D2201 Preservation DataStore Interface D2301 Report on OAIS Access Model D2101B Associated draft reports of knowledge management architecture and tools D2103 Prototype of Descriptive Information-related KM Services D2103 Prototype of Descriptive Information-related KM Services: D2201 Preservation Datastores Interface D2201 Updated Preservation DataStores Interface: D2203 Prototype enhanced version of Preservation DataStores D2302 Accompanying document of prototype of OAIS-based access, Ontology-based DRM services and DAMS services D2303 Prototypes of Authenticity tools and OAIS-based infomration browsing k D3101 Report on Framework Architecture

	D3201 CASPAR Integration with External Systems D3301 Report on Technical testing and integration D4101 User requirements and scenario specifications D4102 Integrated report of R&D activities on the Cultural, Performing Arts and Science data Preservation Activities D4103 REPORT ON EVALUATION CRITERIA D4104 CASPAR VALIDATION/EVALUATION REPORT D4105 CASPAR Draft Testbed Implementation Plan D5101 CASPAR Dissemination and Use Plan D5301 CASPAR Training Plan D5302 CASPAR Overview and Overall Assessment of the Training Activities
Relevance	As CASPAR aimed to investigate not just the bits of digital objects, but also the information and knowledge preservation then it belongs to the interest group for this research. It is also very important that the CASPAR project brought together a consortium covering extensive scientific, cultural and creative expertise in the field of information preservation. The results of the project will also give a very good overview of issues related to the representation information. Therefore it is most reasonable to investigate that project more closely when finding answers to the research questions of this dissertation. Especially valuable is the concept of profiling a Designated Community and formalizing intelligibility by using modules and dependencies. This concept has been explained in deliverables D2101 (Giaretta 2008), D2101B (Tzitzika et al. 2008, 1a) and D2102 (Tzitzika et al. 2008, 1b) of the CASPAR project. Therefore, these deliverables will be the most relevant of CASPAR deliverables for this research. As profiling designated communities is an important part in digital knowledge preservation area, the topic will be further investigated in the light of article II.
Note	All information represented here (except Relevance) has been harvested from the web page of this project or from the CORDIS system* (http://cordis.europa.eu/).

3.1.1.6. E-ARK

Name	European Archival Records and Knowledge Preservation (E-ARK)
Web page	http://www.eark-project.com/
Funded by	European Commission under its ICT Policy Support Programme (PSP) within its Competitiveness and Innovation Framework Programme (CIP)
Duration	From 2014 to 2017
Partners	Coordinator • University of Brighton Partners • AIT Austrian Institute of Technology GmbH (AIT) • Archives of the Republic of Slovenia – Arhiv Republike Slovenije: Ministry of Culture – Ministrstvo za kulturo (MC ARS)

	• The DLM Forum Foundation (DLM) • Statens Arkiver: The Danish National Archives (DNA) • The Digital Preservation Coalition (DPC) • Universität zu Köln (UCO) • Instituto Superior Técnico (IST) • The National Archives of Hungary (NAH) • Rahvusarhiiv: The National Archives of Estonia (NAE) • Kulturdepartementet: Ministry of Culture: National Archival Services of Norway (NAN) • ES Solutions (ESS) • Magenta (MAG) • KEEP SOLUTIONS, LDA (KEEPS) • Agência para a Modernização Administrativa IP – Agency for Public Services Reform, IP (AMA) • Ministerio de Hacienda y Administraciones Pública (MHAP)
Description	In co-operation with commercial systems providers, E-ARK will create and pilot a pan-European methodology for electronic document archiving, synthesising existing national and international best practices, that will keep records and databases authentic and usable over time. Our objective is to provide a single, scalable, robust approach capable of meeting the needs of diverse organisations, public and private, large and small, and able to support complex data types. E-ARK will demonstrate the potential benefits for public administrations, public agencies, public services, citizens and business by providing simple, efficient access to the workflows for the three main activities of an archive - acquiring, preserving and enabling re-use of information. The practices developed within the project will reduce the risk of information loss due to unsuitable approaches to keeping and archiving of records. The project will be public facing, providing a fully operational archival service, and access to information for its users. The project results will be generic and scalable in order to build an archival infrastructure across the EU and in environments where different legal systems and records management traditions apply. E-ARK will provide new types of access for business users. E-ARK will pilot an end-to-end OAIS-compliant e-archival service covering ingest, vendor-neutral archiving, and reuse of structured and unstructured data, thus covering both databases and records, addressing the needs of data subjects, owners and users. The pilot and methodology will also focus on the essential pre-ingest phase of data export and normalisation in source systems. The pilot will integrate tools currently in use in partner organisations, and provide a framework for providers of these and similar tools ensuring compatibility and interoperability. To sustain the outputs of our project, project partner The DLM Forum, comprising 22 national archives and associated commercial and technical providers, is well placed to ensure these. Using the open Apache licensing model, commercial suppliers will be able to incorporate the project outputs (particularly the open interfaces for pre-ingest, ingest, archival, access and re-use) into their own systems, enhancing their

	longevity. National archives running E-ARK pilot instances will serve as exemplars for others wanting to adopt up the new e-archiving open system. In addition, project partner, The Digital Preservation Coalition will promote best practices in this area, as will our dedicated government institution partners.
Progress	D2.1 General pilot model and use case definition D2.2 Legal Issues Report for pilot D2.3 Detailed pilot Requirements D2.4 Pilot Documentation D2.5 Recommended practices and final public report on pilots D3.1 Report on available best practices D3.2 E-ARK SIP draft Specification D3.3 E-ARK SIP pilot specification (revision of D3.2) D3.4 Records export, transfer and ingest recommendations and SIP Creation Tools D4.1 Report on available formats and restrictions D4.2 E-ARK AIP draft Specification D4.3 E-ARK AIP pilot specification (revision of D4.2) D4.4 Final version of SIP-AIP conversion component D5.1 GAP report between requirements for access and current access solutions D5.2 E-ARK DIP draft Specification D5.3 E-ARK DIP pilot specification (revision of D5.2) D5.4 Search, Access and Display Interfaces D6.1 Faceted Query Interface and API D6.2 Integrated Platform Reference Implementation D6.3 Data Mining Showcase D7.1 A Maturity Model for Information Governance – initial version D7.2 Initial Assessment and Evaluation D7.3 Prototype of the Knowledge Centre Service D7.4 The Knowledge Centre Service D7.5 A Maturity Model for Information Governance – final version D7.6 Final Assessment and Evaluation D8.1.1 Annual Dissemination Strategy (Year 1) D8.1.2 Annual Dissemination Strategy (Year 2) (including updated Stakeholder Analysis and ExploitationStrategy) D8.1.3 Annual Dissemination Strategy (Year 3) (including Updated Stakeholder Analysis and Exploitation Strategy) D8.2 Closing International Conference D8.3 End-Project Long-Term Impact Analysis of Dissemination
Relevance	The E-ARK project addresses several important issues of digital preservation of knowledge. Most important are pre-ingest and ingest activities (including the specification of suitable format for a submission information package). The author of this dissertation is also the work package lead of WP3 “Transfer of records to archives” in E-ARK project which eases the research process as a significant amount of specific knowledge is available to use.

	The project has more than 15 partners from different organisations (national archives, research organisations, vendors, universities etc.). One of the partners is the DLM Forum which brings to the E-ARK project skills and knowledge from 22 national archives. Most relevant deliverables from the perspective of this research are: D3.1 Report on available best practices (Kärberg et al. 2014). This deliverable introduces best practices in digital archiving and pays special attention to records export, archival ingest workflows, submission information package formats and ingest of digital objects and their metadata. This report is significant as it helps to understand how memory institutions and other stakeholders deal with digital objects, what is common and what is not. It also reflects that there is no common format for submission information packages (SIPs) or even in some cases clear understanding of what a SIP is, although the term is well described in the OAIS model. D3.2 E-ARK SIP draft Specification (Kärberg et al. 2015). This deliverable describes a draft SIP specification for the E-ARK project and provides initial input for the technical implementations of E-ARK ingest tools. As the format is based on available best practices, it is a good candidate for a common central format for SIPs all over the world. D3.3 E-ARK SIP pilot specification (revision of D3.2) (Kärberg et al. 2016). This deliverable extends the deliverable D3.2 by providing a revised version of its content, adding more details relevant for tool development and implementation, and describing specific profiles for the transfer of relational databases, electronic records management systems (ERMS) and simple file system based records (SFSB). As the author of this thesis was also actively involved in the writing process of the before mentioned documents, then there was good alignment of the relevant outcomes of the E-ARK project and this thesis.
Note	All information represented here (except Relevance) has been harvested from the web page of this project or from the CORDIS system* (http://cordis.europa.eu/).

3.1.1.7. ENSURE

Name	Enabling kNowledge Sustainability Usability and Recovery for Economic (ENSURE)
Web page	http://www.prelida.eu/ensure-enabling-knowledge-sustainability-usability-and-recovery-economic-value http://www.ensure-fp7.eu/ https://ec.europa.eu/research/environment/pdf/project_summaries/fp7/natural_hazards/ensure.pdf (Objectives)
Funded by	European Commission, Seventh Framework (FP7) ICT Programme
Duration	From 2011 to 2014
Partners	Coordinator • IBM Israel – Science and Technology Ltd (Israel)

	Participants • Custodix Nv (Belgium) • Jrc Capital Management Consultancy & Research Gmbh (Germany) • Fraunhofer-Gesellschaft Zur Foerderung Der Angewandten Forschung E.V (Germany) • Atos Spain Sa (Spain) • Fundacion Para El Fomento De La Investigacion Sanitaria Y Biomedica Dela Comunitat Valenciana (Spain) • Maccabi Healthcare Services (Israel) • Philips Electronics Nederland B.V. (Netherlands) • Universidade Do Porto (Portugal) • Lulea Tekniska Universitet (Sweden) • Tessella Ltd (United Kingdom) • Cranfield University (United Kingdom) • Science And Technology Facilities Council (United Kingdom)
Description	The overall objective of ENSURE was to develop a new methodological framework for Integrated Multi-Scale Vulnerability Assessment. More precisely, ENSURE aimed at achieving the following objectives: • to improve the understanding of the articulated nature of the concept of vulnerability at different spatial scales; • to analyze the relationship between the concept of vulnerability and other concepts such as “risk”, “damage”, “exposure”, “resilience” and “adaptation”; • to develop the integration and connection of different types of vulnerability; • to investigate the temporal and spatial variability of the relations between different types of vulnerability and different types of damage; • to propose new, and improve existing vulnerability assessment models and parameters; • to develop an on-line course for students, young researchers and public administration staff in the field of vulnerability assessment.
Progress	Information is not available.
Relevance	The project explored issues specific to digital preservation in the context of the industry and services sector. As the project addressed mainly the long-term usability of commercially relevant data produced or controlled by organisations, then it does not belong to the scope of this research.
Note	All information represented here (except Relevance) has been harvested from the web page of this project or from the CORDIS system* (http://cordis.europa.eu/).

3.1.1.8. ForgetIT

Name	ForgetIT. Concise Preservation by combining Managed Forgetting and Contextualized Remembering.
Web page	http://www.forgetit-project.eu/en/
Funded by	The project is funded by the European Commission within the 7th Framework Programme under the objective "Digital Preservation"
Duration	From 2013 to (2016)
Partners	• Gottfried Wilhelm Leibniz Universität Hannover • Centre for Research and Technology Hellas • dkd Internet Service GmbH • Deutsches Forschungszentrum für Künstliche Intelligenz GmbH • EURIX Srl • IBM Israel - Science and Technology Ltd • Luleå Tekniska Universitet • The Chancellor, Masters and Scholars of the University of Oxford • The University of Edinburgh • The University of Sheffield • Turk Telekomunikasyon AS
Description	While preservation of digital content is now well established in memory institutions such as national libraries and archives, it is still in its infancy in most other organizations, and even more so for personal content. ForgetIT combines three new concepts to ease the adoption of preservation in the personal and organizational context, each overcoming major obstacles: • First, Managed Forgetting models resource selection as a function of attention and significance dynamics. It is inspired by the important role of forgetting in human memory and focuses on characteristic signals of reduction in salience. For this purpose it relies on multi-faceted information assessment and offers customizable preservation options such as full preservation, removing of redundancy, resource condensation, and also complete digital forgetting. • Secondly, Synergetic Preservation crosses the chasm that exists between active information use and preservation management by making intelligent preservation processes an integral part of the content lifecycle in information management and by developing solutions for smooth bi-directional transitions. • Thirdly, Contextualized Remembering targets keeping preserved content meaningful and useful. It will be based on a process of dynamic evolution-aware contextualization, which combines context extraction and packaging with evolution detection and intelligent re-contextualization. To achieve these goals ForgetIT brings together an interdisciplinary team of experts in preservation, information management, information extraction, multimedia analysis, personal information management, storage computing, and cloud computing, as well as in cognitive psychology, law, and economics, who together will develop the innovative methods for realizing the ForgetIT approach.

	The main expected outcomes are the flexible Preserve-or-Forget Framework for intelligent preservation management and, on top of it, two application pilots: one for personal preservation focussing on multimedia coverage of personal events and one for organizational preservation targeted at smooth preservation in organizational content management. The ForgetIT project is expected to have different forms of socio-economic impact: a) The ForgetIT technology increases organizational productivity by enabling more concise forms of organizational memory, thus reducing the risk of work duplication and easing knowledge digestion; b) In addition, the project creates new economic opportunities for actors in technology development and consultation by opening up new application domains for preservation technology such as personal preservation as a service; c) The ForgetIT technology is an important building block for managing and preserving new bottom up forms of community memory and cultural history; and d) the managed forgetting approach provides the first step towards a promising alternative to the prevailing "keep it all" approach in our digital society.
Progress	D11.1 Project Presentation, Project Brochure & Project Web Site D3.1 Report on Foundations of Managed Forgetting D4.1 Information Analysis, Consolidation and Concentration for preservation: State of the Art and Approach D5.1 Foundations of Synergetic Preservation D6.1 State of the Art and Approach for Contextualization D7.1 Foundation of computational storage services D9.1 Application Use Cases & Requirements Document D11.2 Dissemination Plan D2.1 Evaluation Plan (Updated Version) D8.1 Integration Plan and Architectural Approach (Updated Version) D2.2 Foundations of Forgetting and Remembering - Preliminary Report D3.2 Components for Managed Forgetting – First Release D4.2 Information analysis, consolidation and concentration techniques, and evaluation - First release D5.2 Workflow model and prototype for transition between active system and AIS – First release D6.2 Contextualization Tools - First Release D9.2 Personal Preservation Mockups (Annex with Mockups) D8.2 The Preserve-or-Forget Reference Model – Initial Model D8.3 The Preserve-or-Forget Framework – First Release D2.3 Foundations of Forgetting and Remembering: Intermediary Report D3.3 Strategies and Components for Managed Forgetting – Second Release D4.3 Information analysis, consolidation and concentration techniques, and evaluation - Second release D5.3 Workflow model and prototype for transition between active system and AIS – Second release D6.3 Contextualisation Tools - Second Release: Updates to the Context Modelling Framework and Modules D9.3 Personal Preservation Pilot I: Concise Preserving Personal Desktop

Relevance	The project was focused on digital preservation issues of organisational and personal knowledge. As the project started rather late (2013) compared to the time schedule of this research, and as this project is mainly focused on organisational (excluding memory institutions such as national libraries and archives) and personal knowledge preservation, then it will be not further investigated in the scope of this research.
Note	All information represented here (except Relevance) has been harvested from the web page of this project or from the CORDIS system* (http://cordis.europa.eu/).

3.1.1.9. KEEP

Name	Keeping Emulation Environments Portable (KEEP)
Web page	http://www.keep-project.eu/ezpub2/index.php http://www.keep-project.eu/downloads/training/03overview.pdf
Funded by	European Commission, Seventh Framework (FP7) ICT Programme
Duration	From 2009 to 2012
Partners	Coordinator • Bibliotheque Nationale De France (France) Participants • Deutsche Nationalbibliothek (Germany) • Foerderverein Fuer Jugend Und Sozialarbeit E.V. (Germany) • Joguin (France) • Koninklijke Bibliotheek (Netherlands) • European Games Developer Federation Ekonomisk Forening (Sweden) • Tessella Ltd (United Kingdom) • University Of Portsmouth Higher Education Corporation (United Kingdom)
Description	KEEP developed an Emulation Access Platform enabling the accurate rendering of objects produced on obsolete machines, so that they can be securely accessed in the long term. The project was thus addressing the problems related to transferring digital objects stored on outdated computer media (such as floppy discs) to current storage devices in order to reduce the risk of data loss resulting from technical obsolescence. The KEEP emulation tools are flexible enough to handle both static and dynamic digital objects designed for a wide variety of computer systems: text, sound, and image files; multimedia documents, websites, databases, videogames etc. Research involved the specification of file formats and the production of transfer tools exploited within the framework, and the consideration of possible legal and technical issues. Being aware of the fact that emulation software itself is prone to digital obsolescence, the KEEP consortium planned to create a portable platform that can be run on any possible device and be implemented to future computer systems. The emulation framework was distributed as open-source software so that the emulation community as a whole can contribute to its further development.

	In addition to producing a software package, the project improved understanding about how to integrate emulation-based solutions with an operational electronic deposit system. Existing metadata models were researched and guidelines developed for mapping digital objects to emulated manifestations. Through this work, KEEP aimed at creating a foundation for the next generation of permanent access strategies based on emulation. Although primarily aimed at stakeholders in cultural heritage, such as memory institutions and (computer games) museums, the Emulation Access Platform is also able to serve the needs of a wider range of organisations and individuals.
Progress	D7.2 Annual Report (Publishable summary) Year 1 D7.2 Annual Report (Publishable summary) Year 2 D1.2a Requirement specification for the Transfer Tools Framework D1.2b Conceptual Design Document for the Transfer Tools Framework D2.2 Requirements & design documents for the Emulation Framework D2.3 Prototype of Emulation Framework (cover) D2.6 Layman's Guide to KEEP Legal Studies D3.1 Preliminary document analyzing and summarizing metadata standards and issues across Europe D3.2a Document describing metadata for the specified range of digital objects D3.2b Documents describing meta-data for the specified range of digital objects, as well as requirements and design for the browsing system and user interface of the Emulation Framework
Relevance	The project was investigating various aspect of emulation, in theory and in practice, and has prompted many organisations to actually use emulation for the first time. It also provided tools etc. that have led to Emulation as a Service, used now by several memory institutions. Although enabling access to and use of digital objects stored on outdated computer media is important, the main focus of this project is still out of scope of the research carried out in this dissertation. Therefore, the project will not be further investigated.
Note	All information represented here (except Relevance) has been harvested from the web page of this project or from the CORDIS system* (http://cordis.europa.eu/).

3.1.1.10.LIWA

Name	Living Web Archives (LIWA)
Web page	http://liwa-project.eu/
Funded by	European Commission, Seventh Framework (FP7) ICT Programme
Duration	From 2008 to 2011
Partners	Coordinator • Gottfried Wilhelm Leibniz Universitaet Hannover (Germany) Participants • Narodni Knihovna Ceske Republiky (Czech Republic) • Moravska Zemska Knihovna V Brne (Czech Republic)

	• Max Planck Gesellschaft Zur Foerderung Der Wissenschaften E.V. (Germany) • Stichting Internet Memory Foundation (France) • Magyar Tudomanyos Akademia Szamitastechnikai Es Automatizalasi Kutato Intezet (Hungary) • Stichting Nederlands Instituut Voor Beeld En Geluid (Netherlands) • Hanzo Archives Limited (United Kingdom)
Description	The interest in Web content preservation is strongly growing, not only in traditional library and archival organisations, but also in sectors such as industry and services. But the typical characteristics of Web content – variety of formats, high dynamics, volatility, interactivity and context-dependency – make adequate Web archiving a particular challenge. With the LiWA project, Web archiving has been established as a new topic for scientific research and development within the digital preservation domain. At the centre of the project was the concept of 'Living Web Archives', as opposed to the current practice of producing periodic snapshots of pages. 'Living' here refers to: • long term interpretability as the archive evolves and adapts over time, • improved archive fidelity and authenticity by filtering out irrelevant information, • captured content from a wide variety of sources. To enhance archive fidelity and authenticity, LiWA has developed and tested new methods based on content interpretation and intelligent pattern detection of traps and Web spam. This allows reducing the amount of fake content and helping prioritise crawls by automatically detecting content of value. To improve the integrity and temporal, structural and semantic coherence of Web archives, some work was dedicated to temporal Web archive construction. This serves the objective to significantly improve content positioning in time and (topic) space and will lay the foundations for fast and effective access to evolving Web content. To facilitate archive interpretability, LiWA applied methods for semantic and terminology extraction, able to detect and handle evolving semantics, interpretations of domain concepts and terminology. This is a contribution to the task of preserving the usefulness, quality, and accessibility of Web archives over time. For validating the LiWA approach, two demonstrator applications have been built on top of the LiWA services. The applications focus on the social Web and on the special challenge of archiving audio-visual content. The potential benefit of this research is twofold: Archiving institutions will be able to automatically archive higher volumes of dynamic and volatile digital content, resulting in a significant increase of preserved digital content. Archive users will benefit from the higher quality of archive content and improved search services.
Progress	The Rich Media Capture Module - a plug-in dedicated to the capture of streaming video content. The Temporal Coherence Analyser – a plug-in dedicated to the analysis of the temporal coherence of the archived Web content.

	The Spam Assessment Interface – a Web service that enables the quality assessment of the archived Web content. The Semantic Analizer – a component dedicated to the detection of terminology evolution. The Web Archive UI Framework – a client-side framework that helps creating User Interface helpers for Web archive browsing.
Relevance	The project was addressing Web Archiving. The tools developed and demonstrated by this project allow capturing web content from a wide variety of sources. Although LiWA delivered two exemplary applications for demonstrating the benefits of advanced Web archiving to interested stakeholders, they still fell out of this selection as they are focusing on audio-visual streams and social web content respectively, but this dissertation seeks solutions for more universal content.
Note	All information represented here (except Relevance) has been harvested from the web page of this project or from the CORDIS system* (http://cordis.europa.eu/).

3.1.1.11. PLANETS

Name	Preservation and Long-term Access through Networked Services (PLANETS)
Web page	http://www.planets-project.eu/ www.statsbiblioteket.dk/forskning/institutionsrelateret-forskning/planets-1/Projektbeskrivelse_PLANETS.pdf
Funded by	European Commission, Sixth Framework (FP6) IST Work Programme
Duration	From 2006 to 2010
Partners	Coordinator • The British Library (United Kingdom) Participants • Oesterreichische Nationalbibliothek (Austria) • Technische Universitaet Wien (Austria) • Ait Austrian Institute of Technology Gmbh (Austria) • Schweizerisches Bundesarchiv (Switzerland) • Universitaet Zu Koeln (Germany) • Albert-Ludwigs-Universitaet Freiburg (Germany) • Det Kongelige Bibliotek, Nationalbibliotek Og Kobenhavns Universitetsbibliotek (Denmark) • Statsbiblioteket (Denmark) • Ibm Nederland N.V. (Netherlands) • Nationaal Archief (Netherlands) • Koninklijke Bibliotheek (Netherlands) • Microsoft Research Limited (United Kingdom) • The Keeper of Public Records And Historic Manuscripts Commissioner (United Kingdom) • Tessella Plc (United Kingdom) • University of Glasgow (United Kingdom)

Description	The primary goal for PLANETS was to build practical services and tools to help ensure long-term access to digital cultural and scientific assets. The project delivered an integrated production environment for the management of digital information preservation, with a special focus on the needs of libraries and archives. The PLANETS environment supports a number of key preservation functions through: • the preservation planning tool PLATO and services that empower organisations to define, evaluate, and execute high-quality and cost-effective preservation plans • methodologies, tools and services for the characterisation of digital objects that can automatically analyse digital objects and establish significant properties • innovative solutions for performing preservation actions and to ensure rendering of the objects and keeping their properties available. In this context, work has been done on the archiving of relational databases, emulation and remote access to emulation services. Integration and automation can be seen as the two prominent features of the Planets environment. The PLANETS Interoperability Framework integrates the deliverables into a downloadable 'click and install' software package. Within this package, there are role-based routines for administrators, preservation experts and business users, enabling organisations to improve decision-making about long term preservation, ensure long-term access to their valued digital content and control the costs of preservation actions. The Planets Project will deliver a sustainable framework to enable long-term preservation of digital content, increasing Europe's ability to ensure access in perpetuity to its digital information. Planets will deliver: • Preservation Planning services that empower organisations to define, evaluate, and execute preservation • Methodologies, tools and services for the Characterisation of digital objects • Innovative solutions for Preservation Actions tools which will transform and emulate obsolete digital assets • An Interoperability Framework to seamlessly integrate tools and services in a distributed service network • A Testbed to provide a consistent and coherent evidence-base for the objective evaluation of different protocols, tools, services and complete preservation plans • A comprehensive Dissemination and Takeup program to ensure vendor adoption and effective user training. The project will enable organisations to improve decision-making about long term preservation, ensure long-term access to their valued digital content and control the costs of preservation actions through increased automation and scaleable infrastructure. Intensive Dissemination and Take-up activities will ensure the widest possible adoption of results in the user community and enable commercial tool and service providers to compete in a new market place for differentiated preservation services and tools.

Progress	PM/1-D9 Updated gender action plan PP/2-D1 Report on policy and strategy models, the result of the first iteration m14 R PU PP/3-D1 Report on usage models for libraries and for archives, the result of the first iteration PP/4-D2 Preservation Planning Service based on tool prototypes for building and managing preservation plans including decision support PP/7-D1 Report on Comparison of PLANETS with OAIS and expected differences PA/2-D1 A model for a generic description of preservation action tools and their context PA/3-D1 Action Registry glossary PA/5-D2 Tools for specific technical environment 1 PC/4-D3 Evaluation report on existing characterisation tools DT/3-D1 Annual Scientific Digest 1 TB/3-D2 Report on the Evaluation of the Effectiveness of the Testbed and its Use TB/3-D6 Testbed Assessment Report
Relevance	The PLANETS project addressed mainly preservation planning, characterisation of digital objects and automation with scalability. It was a huge and influential project – the project delivered a comprehensive framework and a number of practical tools for digital preservation and access. As the project was more focused on the actions that happen inside an archive (effective preservation planning, preservation characterisation techniques, preservation actions, etc.) rather than ingest or pre-ingest then the PLANETS project is not very strongly related to this thesis. Although some ideas introduced in PLANETS (especially in the preservation panning tool PLATO) will influence the definition of archival user communities in the article II.
Note	All information represented here (except Relevance) has been harvested from the web page of this project or from the CORDIS system* (http://cordis.europa.eu/).

3.1.1.12. PRELIDA

Name	PREserving Linked DAta (PRELIDA)
Web page	http://www.prelida.eu/
Funded by	This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration
Duration	From 2013 to 2014
Partners	• APA – Alliance Permanent Access • CNR – Consiglio Nazionale delle Ricerche- Institute of Information Science and Technologies (Italy) • Europeana Foundation • Semantic Technology Institute International • The University of Huddersfield • University of Innsbruck

Description	The objectives of PRELIDA were to: • Raise awareness of existing Digital Preservation solutions in the Linked Data community, and facilitate their uptake to provide support, where possible, for the long-term preservation of Linked Data. • Collect, organise and publish use cases related to the long-term access to Linked Data. The use cases will be contributed by stakeholders that have Linked Data in their long-term business strategies, or by third parties that see Linked Data as a market opportunity. • Create a comprehensive state of the art on the technologies related to Linked Data and Digital Preservation, and set up a Technology/ Research observatory in order to identify the most significant actors in the area of solutions to Linked Data and Digital Preservation challenges. • Bring together internationally renowned scientists and representatives of key stakeholders from both communities, and highlight latest advances in their areas. Facilitate them in exchanging experiences and in discussing latest progress and findings in hard research problems relevant to their areas. Facilitate them in working out specific characteristics of Linked Data that make existing Digital Preservation solutions not fully satisfactory. Then, collectively identify relevant challenges and paths of addressing them in years to come. • Nurture the building of a multidisciplinary research, technology and application community around the domain of preserving Linked Data. • Build new strategic alliances among stakeholders in the Digital Preservation and Linked Data communities. • Influence the design of new research programmes addressing the identified R&D priorities, bringing together Digital Preservation and Linked Data, in the 2020 agenda of the European Commission. • Largely disseminate the findings of the action all over Europe and beyond.
Progress	D2.1 Establishment of the Working Group D2.2 Final report on the opening workshop D2.3 Deployment of the online infrastructure D2.4 Report on the first summer school D2.5 Report on the midterm workshop D2.6 Report on the consolidation and dissemination workshop D2.7 Report on the second summer school D3.1 State of the art D3.2 Consolidated state of the art D4.1 Analysis of the limitations of Digital Preservation solutions for preserving D4.2 First version of roadmap D4.3 Consolidated roadmap D5.1 Dissemination Strategy D5.2 Report on Dissemination Activities
Relevance	As the project aimed to build bridges across the digital preservation and linked data communities (making the linked data community aware of existing outcomes of the digital preservation community) then it is also an important project for this dissertation. Although the dissertation is

	not focused on linked open data, it is still part of the knowledge preservation and will be further investigated in the light of article IV.
Note	All information represented here (except Relevance) has been harvested from the web page of this project or from the CORDIS system* (http://cordis.europa.eu/).

3.1.1.13. PrestoPRIME

Name	PrestoPRIME
Web page	http://www.prestoprime.org/
Funded by	European Commission, Seventh Framework (FP7) ICT Programme
Duration	From 2009 to 2012
Partners	Coordinator • Institut National de l'Audiovisuel (France) Participants • Joanneum Research Forschungsgesellschaft Mbh (Austria) • Osterreichischer Rundfunk (Austria) • Universitaet Innsbruck (Austria) • Highlands Technologies Sas (France) • Ex Libris Ltd. (Israel) • Rai-Radiotelevisione Italiana Spa (Italy) • Eurix Srl (Italy) • Stichting Europeana (Netherlands) • Vereniging Voor Christelijk Hoger Onderwijs Wetenschappelijk Onderzoek En Patientenzorg (Netherlands) • Stichting Nederlands Instituut Voor Beeld En Geluid (Netherlands) • The University of Liverpool (United Kingdom) • University of Southampton (United Kingdom) • British Broadcasting Corporation (United Kingdom)
Description	The project activities were guided by objectives: • to research and develop means of ensuring the permanence of digital audiovisual content in archives, libraries, museums and other collections; • to research and develop means of ensuring the long-term future access to audiovisual content in dynamically changing contexts; \\n • to integrate, evaluate and demonstrate tools and processes for audio-visual digital permanence and access;to establish a European net-worked Competence Centre to gather the knowledge created through the research collaboration and share it with the stakeholder community.
Progress	PP_WP1_D1.0.5_FinalReport_v1.10p PP_WP4_D4.0.6_Proof Of Concept Rights Management System PP_WP4_ID4.0.5b_RightsOntology PP_WP2_D2.1.4_Tools for quantitative comparison of preservation strategies PP_WP4_ID4.0.2e_Annex_A_v1.01.Annex A : Non-blocking service calls PP_WP4_ID4.0.2c_Interface specification of vocabulary alignment and annotation services

	PP_WP4_ID4.0.2b_ Interface specification of metadata conversion and deployment services PP_WP4_ ID4.0.2a Rights Metadata APIs D7.1.5 Audiovisual Digital Preservation Status Report 3 D5.2.2 First Prototype of Open PrestoPRIME Reference Implementation D4.0.2 Metadata and Rights APIs and Service Interfaces D4.2.1 Vocabulary alignment Methodology D7.1.4 Audiovisual Digital Preservation Status Report 2 D2.1.3 AV Data Model : Final Specification D6.3.1 Financial Models and Calculation Mechanisms D2.2.3 Strategy for Use of Preservation Metadata for within a Digital Library with examples of use in audiovisual preservation D2.1.2 Tools for modelling and simulating migration-based preservation D2.2.2 Metadata Models, Interoperability Gaps, and Extensions to Preservation Metadata Standards D5.2.1 Definition and Design of a PrestoPRIME Reference Architecture for the Integration Framework ID3.4.1 Service Level Agreements for Audiovisual Preservation Services ID3.3.1 Media formats, identification methods and implementations for multivalent preservation ID3.2.1: Threats to data integrity from use of large-scale data management environments ID3.1.1 Specification and Design of a Preservation Environment for Audiovisual Content D3.1: Design and Specification of the Audiovisual Preservation Toolkit D6.2.2 European Digital Library implementation guidelines for audiovisual archives D2.3.1 Service-Oriented Models for Audiovisual Content Storage D2.2.1: Preservation Process Modelling (Including a Review of Semantic Process Modelling and Workflow Languages) D2.1.1 Audiovisual preservation strategies, data models and value-chains D7.1.3 Audiovisual Digital Preservation Status Report D5.1.1 Definition of Scenarios PrestoPrime Glossary of Rights
Relevance	The project was addressing long-term preservation of and access to digital audio-visual content by integrating media archives with European on-line digital libraries. Although the project developed a number of technical solutions for managing the preservation content, it is not strictly in the scope of this research, as this research targets a more universal approach to the content – the dissertation deals with research questions which are not audio-visual content specific. The project will not be further investigated.
Note	All information represented here (except Relevance) has been harvested from the web page of this project or from the CORDIS system* (http://cordis.europa.eu/).

3.1.1.14. PROTAGE

Name	PReservation Organizations using Tools in AGent Environments (PROTAGE)
Web page	http://www.ra.ee/protage
Funded by	European Commission, Seventh Framework (FP7) ICT Programme
Duration	From 2007 to 2010
Partners	Coordinator • Riksarkivet (Sweden) Participants • Fraunhofer-Gesellschaft zur Foerderung der Angewandten Forschung E.V (Germany) • Rahvusarhiiv (Estonia) • Easy Innova S.L. (Spain) • Exact Learning Solutions Spa (Italy) • Lulea Tekniska Universitet (Sweden) • University of Bradford (United Kingdom)
Description	PROTAGE addressed the challenges related to the preservation of digital resources of increasing volume and heterogeneity. The solution proposed was to develop tools allowing for more efficiency and self-reliance of preservation processes. For this purpose, PROTAGE researchers explored the value of a promising technology – software agents – for the automation of digital preservation processes. Based on the latest research on digital preservation strategies and on autonomous systems, the project has built and validated flexible and extensible software agents for long-term digital preservation and access that can cooperate with and be integrated in existing and new preservation systems to support various aspects of the digital preservation workflow such as the submission / ingestion of digital material, monitoring of preservation systems and transfer between repositories. Tools developed by the PROTAGE project will: • enable content producers to create and publish in a preservation-compatible manner, • provide digital repositories with means of further automating the preservation processes, • facilitate seamless interoperation between content providers, libraries and archives, and end-users throughout Europe. Targeted end users are curators and digital content creators, including individuals managing their own digital collections. PROTAGE has used archive and library materials from the project partners for initial system and user tests and material from external stakeholders for further validation. The Swedish Centre of Competence for Long-term Preservation ensures the availability of results to a wider community of memory institutions. The industrial partners have the possibility to use the research results to develop commercial solutions. Objectives: • Research the potential of software agent ecosystems to support the automation of digital preservation tasks;

	• Demonstrate the technical feasibility of such a system; • Analyse implementation in different organisational environments; • Explore possible integration with other digital preservation environments; • Explore synergies with other RTD activities in digital preservation.
Progress	D1.1 State-of-the-art, Stakeholder Needs, Application Scenarios D1.3 PROTAGE Methodology Handbook D7.2 Dissemination Master Plan Agents: A Quick State of the Art Agents: A Deeper State-of-the-Art Analysis – Part I Potential of Agents in Digital Preservation.
Relevance	The project built and validated software agents for long-term digital preservation and access that can be integrated into existing and new preservation systems. As the project looked for ways to coordinate the pre-ingest phase more effectively then the results of the PROTAGE project will be respectfully observed. A more detailed look to the documentation (NAE 2010) will be taken in the preparation phase of the article I of this research.
Note	All information represented here (except Relevance) has been harvested from the web page of this project or from the CORDIS system* (http://cordis.europa.eu/).

3.1.1.15. SCAPE

Name	Scalable Preservation Environments (SCAPE)
Web page	www.scape-project.eu/
Funded by	European Commission, Seventh Framework (FP7) ICT Programme
Duration	From 2011 to 2014
Partners	Coordinator • AIT Austrian Institute of Technology GmbH (Austria) Participants • Technische Universitaet Wien (Austria) • Oesterreichische Nationalbibliothek (Austria) • Vysoke Uceni Technicke V Brne (Czech Republic) • Fachinformationszentrum Karlsruhe Gesellschaft Fur Wissenschaftlich-Technische Information Gmbh (Germany) • Technische Universitaet Berlin (Germany) • Statsbiblioteket (Denmark) • Universite Pierre Et Marie Curie – Paris 6 (France) • Ex Libris Ltd. (Israel) • Stichting Internet Memory Foundation (Netherlands) • Koninklijke Bibliotheek (Netherlands) • Instytut Chemii Bioorganicznej Polskiej Akademii Nauk (Poland) • Wielkopolskie Centrum Pulmonologiii Torakochirurgii Im. Eugenii I Janusza Zeylandow (Poland) • Keep Solutions Lda (Portugal) • Universitatea De Vest Din Timisoara (Romania)

	• Microsoft Research Limited (United Kingdom) • The University Of Manchester (United Kingdom) • The British Library Board (United Kingdom) • Open Planets Foundation (United Kingdom) • Science and Technology Facilities Council (United Kingdom) • Centre National De La Recherche Scientifique (France)
Description	The SCAPE project enhanced the state of the art of digital preservation in three ways: by developing infrastructure and tools for scalable preservation actions; by providing a framework for automated, quality-assured preservation workflows and by integrating these components with a policy-based preservation planning and watch system. These concrete project results were validated within three large-scale testbeds from diverse application areas: Digital Repositories from the library community, Web Content from the web archiving community, and Research Data Sets from the scientific community. Each Testbed has been selected because it highlights unique challenges. SCAPE developed scalable services for planning and execution of institutional preservation strategies on an open source platform that orchestrates semi-automated workflows for large-scale, heterogeneous collections of complex digital objects. These services • Identify requirements for preserving all or parts of a repository through characterisation and trend analysis • Define responses to those needs using formal descriptions of preservation policies and preservation plans • Allow a high degree of automation, virtualization of tools, and scalable processing • Monitor the quality of preservation processes. The SCAPE consortium brought together experts from memory institutions, data centres, research labs, universities, and industrial firms in order to research and develop scalable preservation systems that can be practically deployed within the project lifetime. SCAPE was dedicated toward producing open source software solutions available to the entire digital preservation community. The project results are curated and further exploited by the newly founded Open Planets Foundation. Project results are being exploited by an SME and research institutions within the consortium catering to the preservation community and by two industrial IT partners.
Progress	D23.2 Final report on preservation scenarios deployed and integrated with data centres D18.2 SCAPE final evaluation and methodology report D12.3 Final version of the simulation environment D11.3 QA Workflow, Release 3 + Release Report D10.3 Final version of action components D9.3 Characterisation technology, Release 3 + release report D8.2 Reference implementation of DOR D3.2 Report on future research challenges D19.2 Final Demonstration Report D14.2 Final version of automated policy-aware planning component

	D11.4 Knowledge base of format-specific preservation risks D5.3 Application Integration and Provisioning D22.2 Final Sustainability Plan D23.1 Methodology: contextual information and semantic information for video processing D22.1 Business case template D17.2 Research data sets executable workflows for large-scale execution D16.2 LSDR executable workflows for large-scale execution D15.2 Web content executable workflows for large-scale execution D8.3 Initial anonymized ingestion prototype D7.3 Design and implementation of the preservation component catalogue D4.3 Final Data Center Deployments D12.2 Final version of the Preservation Watch component D6.3 Optimization of Preservation Processes D2.3 Technical Architecture Report v2 D20.6 Final best practice guidelines and recommendations D13.2 Catalogue of preservation policy elements D8.1 Recommendations for Preservation-aware Digital Object Model D7.2 Workflow modelling environment D4.2 Final Platform Release D6.2 Demonstrator and report on workflow compilation and parallel execution D13.1 Final Version of Policy Specification Model D11.2 Quality Assurance Workflow, Release 2 + Release Report D9.2 Characterisation technology, Release 2 + release report D5.2 Job submission language and interface D6.1 Report on the feasibility of parallelising preservation processes D4.1 Architecture Design (First Version) D18.1 First evaluation report D10.2 Gap analysis on action services tools and SCAPE platform and testbeds requirements D3.1 Open Research Challenges and Research Roadmap for SCAPE D2.2 Technical Architecture Report v1 D7.1 Design of provenance component D21.1 Training Plan D17.1 Research datasets executable workflows for experimental execution D16.1 LSDR Executable Workflows for Experimental Execution D15.1 Web content executable workflows for experimental execution D11.1 Quality Assurance Workflow, Release 1 & Release Report D9.1 Characterisation technology, Release 1 & release report D5.1 Guidelines for deploying preservation tools and environments D12.1 Preservation Watch Component Architecture D14.1 Report on Decision Factors and their Influence on Planning D20.1 Communication and Dissemination Plan D10.1 Identification and selection of large-scale migration tools and services D2.1 Technical Implementation Guidelines

Relevance	The project primarily addressed the scalability issue of digital preservation. SCAPE delivered scalability improvements on some existing tools (i.e. Plato), new scalable tools (i.e. Hawarp) and tools for quality control (i.e. Matchbox). As the research questions of this dissertation are more specific to digital preservation of knowledge then the SCAPE project will not be further investigated.
Note	All information represented here (except Relevance) has been harvested from the web page of this project or from the CORDIS system* (http://cordis.europa.eu/).

3.1.1.16. SHAMAN

Name	Sustaining Heritage Access through Multivalent Archiving (SHAMAN)
Web page	http://www.shaman-ip.eu/
Funded by	European Commission, Seventh Framework (FP7) ICT Programme
Duration	From 2007 to 2011
Partners	Coordinator • Inmark Estudios Y Estrategias S.A. (Spain) Participants • Incontec Gmbh (Germany) • Georg-August-Universitaet Goettingen Stiftung Oeffentlichen Rechts (Germany) • Fernuniversitaet In Hagen (Germany) • Deutsche Nationalbibliothek (Germany) • Otto-von-Guericke-Universitaet Magdeburg (Germany) • Globit-Globale Informationstechnik Gmbh (Germany) • Xerox Sas (France) • Philips Consumer Lifestyle B.V. (Netherlands) • Inesc Id – Instituto De Engenharia De Sistemas E Computadores, Investigacao E Desenvolvimento Em Lisboa (Portugal) • Hoegskolan I Boras (Sweden) • Industrious Media Limited (United Kingdom) • The University of Liverpool (United Kingdom) • The University of Glasgow (United Kingdom) • University of Strathclyde (United Kingdom) • The Board of Trustees of The University Of Illinois (United States)
Description	The SHAMAN Integrated Project aimed at developing a new framework for long-term digital preservation (more than one century) by exploring the potential of recent developments in the areas of GRID computing, federated digital library architectures, multivalent emulation and semantic representation and annotation. The researchers' vision was: "For the longer term, SHAMAN will develop radically new approaches to Digital Preservation, such as those inspired by human capacity to deal with information and knowledge, providing a sound basis and instruments for unleashing the potential of advanced ICT to automatically act on high volumes and dynamic and

	volatile digital content, guaranteeing its preservation, keeping track of its evolving semantics and usage context and safeguarding its integrity, authenticity and long term accessibility over time." The project planned to deliver a set of integrated tools supporting the various aspects of the preservation process: analysis/characterisation, ingestion, management, access and reuse. Work included trials and validation of the tools in three application domains dealing with different types of objects: scientific publishing and government archives, industrial design and engineering (e.g. CAD), and e-science resources. SHAMAN's dissemination and exploitation plans aimed at actively fostering outreach and take-up of results and will be tailored according to the specific needs of the scientific / academic world and of industry users. SHAMAN's work was coordinated with other digital preservation projects and initiatives at national and international level.
Progress	Information not available.
Relevance	The project developed tools for analysing, ingesting, managing, accessing and reusing information objects and data across memory institutions. It also explored the ways of using GRID computing in digital preservation. As there is not enough public information available about the outcomes of this project then it cannot be further investigated.
Note	All information represented here (except Relevance) has been harvested from the web page of this project or from the CORDIS system* (http://cordis.europa.eu/).

3.1.1.17. TIMBUS

Name	Timeless Business Processes and Services (TIMBUS)
Web page	http://timbusproject.net/
Funded by	European Commission, Seventh Framework (FP7) ICT Programme
Duration	From 2011 to 2014
Partners	Coordinator • Sap Ag (Germany) Participants • Verein zur Forderung der It-Sicherheit in Osterreich (Austria) • Sqs Software Quality Systems Ag (Germany) • Westfaelische Wilhelms-Universitaet Muenster (Germany) • Karlsruher Institut Fuer Technologie (Germany) • Intel Research And Innovation Ireland Limited (Ireland) • Inesc Id – Instituto De Engenhariade Sistemas E Computadores, Investigacao E Desenvolvimento Em Lisboa (Portugal) • Laboratorio Nacional De Engenharia Civil (Portugal) • Caixa Magica Software Lda (Portugal) • Digital Preservation Coalition Limited By Guarantee *Dpc (United Kingdom)
Description	A primary motivation for TIMBUS was the declining use of centralised in-house business processes. Based on the concept of software as a service

	and on Internet services, business processes are increasingly supported by systems offered by different providers and located in various geographical locations. These are not only continuously changing and evolving but might even be discontinued if services or service providers disappear, leaving partially complete business processes behind. Therefore, TIMBUS researchers endeavoured to enlarge the understanding of digital preservation to include the set of activities, processes and tools that ensure continued access to services and software. This enlarged understanding brought digital preservation into the domain of business continuity management. This approach allowed for reproducing the execution context in which information can be accessed, properly rendered, validated and transformed into knowledge. One of the fundamental requirements for achieving this was to preserve the functional and non-functional specifications of services and software, along with their dependencies. The TIMBUS consortium showed a strong participation of the IT industry, combined with industrial and academic research labs.
Progress	D2.3 M36 Exploitation Plan Iter. 3 – redacted for public D3.2 M12 icon Dissemination and Standardisation Plan and Activities Iter1 D3.3 M24 Dissemination/ Standardisation Plan & Activities Iter. 2 D3.4 M 24 Training Material & Programme Iter. 1 D4.2 M12 Dependency Models Iter1 D4.3 M24 Dependency Models Iter. 2 D4.5 M12 Business Process Contexts D4.6 M24 Use Case Specific DP & Holistic Escrow D4.7 M36 Validation of DP'ed Business Processes and Verification of Redeployed Business Processes D4.9 M36 Refined Business Process Contexts D5.1 M12 Architecture for Intelligent ERM D5.3 M24 Architecture for VSRI D5.4 M18 Refined Architecture for intelligent ERM D6.6 M36 Business Process Preservation Test Bed D7.1 M18 Process and methods for digitally preservable services D8.1 M12 Use Case Definition and Requirements - Civil Enginieering Infrastructures D9.1 M12 Use Case Definition and Requirements – High Energy Physics
Relevance	The project explored issues specific to digital preservation of enterprise business processes. As the project addressed mainly the long-term usability of commercially relevant data processes (aligning with the enterprise risk management and business continuity management) then it does not belong to the scope of this research as this thesis is focused on the knowledge preservation issues related to memory institutions.
Note	All information represented here (except Relevance) has been harvested from the web page of this project or from the CORDIS system* (http://cordis.europa.eu/).

3.1.1.18. Wf4Ever

Name	Advanced Workflow Preservation Technologies for Enhanced Science (Wf4Ever)
Web page	http://www.wf4ever-project.org/ http://cordis.europa.eu/project/rcn/97462_en.html
Funded by	European Commission, Seventh Framework (FP7) ICT Programme
Duration	From 2010 to 2013
Partners	Coordinator • Intelligent Software Components S.A. (Spain) Participants • Universidad Politecnica de Madrid (Spain) • Agencia Estatal Consejo Superior de Investigaciones Cientificas (Spain) • Academisch Ziekenhuis Leiden – Leids Universitair Medisch Centrum (Netherlands) • Instytut Chemii Bioorganicznej Pan (Poland) • The University of Manchester (United Kingdom) • The Chancellor, Masters And Scholars of The University of Oxford (United Kingdom)
Description	Wf4Ever developed new models, techniques and tools for the preservation of scientific workflows, including the novel definition of a Research Object, which packages workflow descriptions, the provenance of their executions, and linked to all the related resources upon which they depend. Such models also included models for repeatability and reproducibility, and models for workflow abstraction, to facilitate workflow classification and indexing, comparison, and similarity detection between pairs of existing workflows in the library. Wf4Ever also developed strategies for sharing and reusing workflows or workflow fragments and patterns, including mechanisms for personalised workflow recommendation based on workflow descriptions, users' collective behaviour, and social information. Finally, Wf4Ever proposed methods and tools to proactively preserve and inspect workflow integrity and authenticity through the evaluation of workflow information quality, based on the provenance of workflows and their research objects and described in new vocabularies for the representation of the provenance of research objects in digital preservation systems. Wf4Ever developed a software architecture and reference implementation for the preservation of scientific workflows, which extended one of the most widely deployed scientific workflow sharing infrastructures (myExperiment) with preservation capabilities that consider the complexity of scientific workflows and their related objects. This software system leveraged the advances done on workflow lifecycle management, collaboration and sharing support, and integrity and authenticity maintenance. Wf4Ever evaluated in two workflow-intensive use cases in the domains of Astronomomy and Genomics.
Progress	D1.1 Setup of software development technologies D1.2 Wf4Ever Sandbox v1 D1.2v2 Wf4Ever Sandbox – Phase II

	D1.2v3 Wf4Ever Sandbox – Phase III D1.2v4 Wf4Ever Sandbox – Phase IV Final D1.3v1 Wf4Ever Architecture – Phase I D1.3v1 Addendum: Wf4Ever Architecture – Phase I D1.3v2 Wf4Ever Architecture – Phase II D1.4v1 Reference Wf4Ever Implementation – Phase I D1.4v2 Reference Wf4Ever Implementation – Phase II D1.5 Opportunities for applying Wf4Ever architecture and technologies D2.1 Workflow Lifecycle Management Initial Requirements D2.2v1 Design, implementation and deployment of workflow lifecycle management components – Phase I D2.2v2 Design, implementation and deployment of workflow lifecycle management components – Phase II D2.3 Final evaluation report of workflow lifecycle management components D3.1 Workflow Evolution, Sharing and Collaboration Initial Requirements D3.2v1 Design, implementation and deployment of Workflow Evolution, Sharing and Collaboration components D3.2v2 Design, implementation and deployment of Workflow Evolution, Sharing and Collaboration components - Phase II) D3.3 Final evaluation report of Workflow Evolution, Sharing and Collaboration components D4.1 Workflow Integrity and Authenticity Maintenance Initial Requirements D4.2v1 Design, implementation and deployment of Workflow Integrity and Authenticity Maintenance components – Phase I D4.2v2 Design, implementation and deployment of Workflow Integrity and Authenticity Maintenance components – Phase II D4.3 Final evaluation report of the workflow integrity and authenticity maintenance components D5.1 Astronomy Workflow Preservation Requirements D5.2 Creation of an Astrophysical Community of Preserved Workflow Users D5.3v1 Propagation of interdependent quantities in the calculation of luminosities of galaxies D5.3v2 Calculation of Luminosity profiles for a Sample of Galaxies extracted from Catalogues using Isolation Criteria D5.3v3 Astronomy Workflows v3 D6.1 Genomics Workflow Preservation Requirements D6.1 Addendum: Biological User Requirements D6.2 Final Report on the creation of a Community of Users in Genomics D6.3v1 Genome Wide Association Study Workflows v1 D6.3v2 Genomic Wide Association Study Workflows v2 D6.3v3 Genomic Wide Association Study Workflows v3 D7.1 Quality and Risk Contingency Plan D8.1 Wf4Ever web site) D8.2 Dissemination, Transfer and Exploitation D8.4v1 Report on Dissemination Activities v1 D8.4v2 Report on Dissemination Activities v2

Relevance	The project aimed at providing the methods and tools required to ensure the long-term preservation of scientific workflows. As the dissertation deals with more knowledge-oriented research questions (more specific to digital preservation of knowledge and not science-centric) then the project will not be further investigated.
Note	All information represented here (except Relevance) has been harvested from the web page of this project or from the CORDIS system* (http://cordis.europa.eu/).

Appendix B. Local Submission Information Package Specification

Name:
The XML Schema for an archival hierarchy/scheme.
Available at:
http://www.arhiiv.ee/public/Digiarhiiv/UAM/schemas/export/UAM_Eksport_arhiiviskeem_EE_v2.0.xsd

Name:
The XML Schema for an archival records.
Available at:
http://www.arhiiv.ee/public/Digiarhiiv/UAM/schemas/export/UAM_Eksport_arhivaal_EE_v2.0.xsd

Appendix C. International Submission Information Package Specification

Name:
The deliverable D3.3 E-ARK SIP Pilot Specification.
Available at:
http://eark-project.com/resources/project-deliverables/51-d33pilotspec/file

Name:
The deliverable D3.3 E-ARK SMURF (semantically marked up record format).
Available at:
http://eark-project.com/resources/project-deliverables/52-d33smurf/file

CURRICULUM VITAE

Name: Tarvo Kärberg
Date of Birth: 01.06.1984
Citizenship: Estonian
E-mail: karberg@ut.ee
Phone: +372 738 7585

Education

2010 – …	PhD Studies in History, University of Tartu
2006 – 2008	MSc in Information Technology (cum laude), University of Tartu

Work Experience

2014 – ...	European Archival Records and Knowledge Preservation (E-ARK) project, Work Package 3 (Transfer of Records to Archives) Lead
2007 – ...	Rahvusarhiiv (The National Archives of Estonia), Project Manager

Research Interest

Particular interest is in digital preservation of knowledge.

Professional Development

2009 archivist certification

Selection of Publications

I. Digital Preservation of Knowledge in the Public Sector: A Pre-Ingest Tool. (DOI 10.1007/s10502-013-9211-z).
II. Digital Preservation and Knowledge in the Public Archives: For Whom? (DOI 10.1080/23257962.2014.942606).
III. Toward Common Ontologies of Facets of the Archival Access Portal. (DOI 10.1007/978-3-319-03437-9_36).
IV. Transforming users' knowledge into archival knowledge. (DOI: 10.1045/march2016-karberg).

ELULOOKIRJELDUS

Nimi: Tarvo Kärberg
Sünniaeg: 1. juuni, 1984, Eesti
Kodakondsus: Eesti
E-post: karberg@ut.ee
Telefon: +372 738 7585

Hariduskäik:
2010 – … Tartu Ülikool, doktorantuur
2006 – 2008 Tartu Ülikool, MSc (infotehnoloogia, cum laude)

Teenistuskäik:
2014 – ... Tööpaketi 3 (Transfer of Records to Archives) juht projektis European Archival Records and Knowledge Preservation (E-ARK)
2007 – ... Rahvusarhiiv, projektijuht

Uurimisvaldkond:
Teadmuse digitaalse arhiveerimise uurimine.

Erialane enesetäiendus
2009 arhivaari kutseeksami sooritamine

Olulisemad publikatsioonid:
I. Digital Preservation of Knowledge in the Public Sector: A Pre-Ingest Tool. (DOI 10.1007/s10502-013-9211-z).
II. Digital Preservation and Knowledge in the Public Archives: For Whom? (DOI 10.1080/23257962.2014.942606).
III. Toward Common Ontologies of Facets of the Archival Access Portal. (DOI 10.1007/978-3-319-03437-9_36).
IV. Transforming users' knowledge into archival knowledge. (DOI: 10.1045/march2016-karberg).

DISSERTATIONES HISTORIAE UNIVERSITATIS TARTUENSIS

1. **Olaf-Mihkel Klaassen.** Eesti Vabariigi konsulaarpoliitika Aasias ja Aafrikas 1918–1940. Tartu, 1991.
2. **Jüri Linnus.** Maakäsitöölised Eestis 16. sajandist kuni 19. sajandini. Tartu, 1991.
3. **Vahur Made.** Eesti ja rahvasteliit. Tartu, 1999.
4. **Martin Hallik.** Tartu ülikooli õppejõudude ja kasvandike osast humanitaar-orientalistikas (1802–1940). Tartu, 2001.
5. **Anti Selart.** Liivimaa ja Vene 13. sajandil. Uurimus poliitilisest ajaloost. Tartu, 2002.
6. **Aldur Vunk.** Ristisõjad ja palverännakud Eestis 12.–16. sajandil. Uurimus nende iseloomust ja alatüüpidest. Tartu, 2003.
7. **Andres Andresen.** Luterlik territoriaalkirik Eestimaal 1710–1832. Riigivõimu mõju kirikuvalitsemisele, -institutsioonidele ja -õigusele. Tartu, 2004.
8. **Katri Raik.** Eesti- ja Liivimaa kroonikakirjutuse kõrgaeg 16. sajandi teisel poolel ja 17. sajandi alul. Tartu, 2004.
9. **Aigi Rahi-Tamm.** Teise maailmasõja järgsed massirepressioonid Eestis: allikad ja uurimisseis. Tartu, 2004.
10. **Anu Raudsepp.** Ajaloo õpetamise korraldus Eesti NSV eesti õppekeelega üldhariduskoolides 1944–1985. Tartu, 2005.
11. **Lea Leppik.** Tartu ülikooli teenistujate sotsiaalne mobiilsus 1802–1918. Tartu, 2006.
12. **Karin Hiiemaa.** Aafrika retseptsioon eestikeelses trükisõnas (kuni 1917). Tartu, 2006.
13. **Людмила Дубьева.** Историческая наука в Тартуском университете в конце XIX – начале XX вв. Тарту, 2006.
14. **Sirje Tamul.** Eraalgatuslikest stipendiumidest Tartu Ülikoolis 1802–1918. Tartu, 2007.
15. **Marten Seppel.** Näljaabi Liivi- ja Eestimaal 17. sajandist 19. sajandi alguseni. Tartu, 2008.
16. **Mati Kröönström**. Eesti sõjaväe juhtivkoosseis Vabadussõjas 1918–1920. Tartu, 2008.
17. **Märt Läänemets.** *Gaõóavyūha-sūtra* kui ajalooallikas. Tartu, 2009.
18. **Epi Tohvri.** Valgustusideede mõju Tartu arhitektuurikultuurile 19. sajandi alguses. Tartu, 2009.
19. **Indrek Paavle.** Kohaliku halduse sovetiseerimine Eestis 1940–1950. Tartu, 2009.
20. **Aivar Põldvee.** Bengt Gottfried Forselius ja rahvahariduse lätted Eesti- ja Liivimaal. Tartu, 2010.

21. **Vladimir Sazonov.** Die Königstitel und -epitheta in Assyrien, im Hethiterreich und in Nordsyrien (*Ugarit, Emar, Karkemiš*) in der mittelassyrischen Zeit: Strukturelle Gemeinsamkeiten, Unterschiede und gegenseitige Beeinflussung. Tartu, 2010.
22. **Andres Seene.** Eesti sõjaväe ohvitseride ettevalmistamise süsteemi kujunemine ja areng 1919–1940. Tartu, 2011.
23. **Piret Õunapuu**. Eesti rahva muuseumi loomine ja väljakujunemine. Tartu, 2011.
24. **Ilmar Rootsi**. Hunt ja inimene: suhted Eestis XVIII sajandi keskpaigast XIX sajandi lõpuni. Tartu, 2011, 282 lk.
25. **Kaja Kumer-Haukanõmm.** Teisest maailmasõjast tingitud Balti pagulaste problemaatika aastatel 1945–1952 Eesti pagulaste näitel. Tartu, 2012, 202 lk.
26. **Meelis Maripuu.** Omavalitsuseta omavalitsused Halduskorraldus Eestis Saksa okupatsiooni ajal 1941–1944. Tartu, 2012, 332 lk.
27. **Kaarel Vanamölder.** Kommunikatsiooniväli Rootsi Läänemere-provintsides 17. sajandi lõpul – Revalsche Post-Zeitung varauusaegse informatsioonikandjana. Tartu, 2012, 221 lk.
28. **Laine Randjärv.** Loovisiksuse roll Eesti laulupeoliikumises aastatel 1940–1980. Tuudur Vettiku ja Roland Laasmäe epistolaarse pärandi põhjal. Tartu, 2013, 305 lk.
29. **Meelis Saueauk.** Nõukogude julgeolekuorganite ja Eestimaa Kommunistliku Partei koostöö Eesti sovetiseerimisel aastatel 1944–1953. Tartu, 2013, 355 lk.
30. **Ivo Juurvee.** Riigisaladuse kaitse Eesti Vabariigis 1918–1940. Tartu, 2013, 282 lk.
31. **Mari Nõmmela.** Stiiliajaloolise ja marksistliku käsitluse konflikt Voldemar Vaga kunstiajalookirjutuses 20. sajandi II poolel. Tartu, 2013, 213 lk.
32. **Kristi Kukk.** Väikerahvuste ajalookäsitluste genees ja narratiivid: Eesti võrdluses teiste Põhjala ja Baltikumi mittedominantsete rahvustega 19. sajandist kuni Teise maailmasõjani. Tartu, 2013, 180 lk.
33. **Helen Rohtmets-Aasa.** Eesti Vabariigi sisserändepoliitika aastatel 1920–1923. Tartu, 2014, 166 lk.
34. **Olev Liivik**. Eesti NSV Ministrite Nõukogu institutsionaalne areng ja kaadrid 1940–1953. Tartu, 2014, 349 lk.
35. **Reigo Lokk**. Sepistades natsiooni: taasiseseisvunud Eesti etnopoliitilised konfliktid. Tartu, 2015, 465 lk.
36. **Tõnis Tatar.** Kolmas tee Eesti NSV kunstis: avangardi ja võimumeelsuse vahel. Tartu, 2015, 295 lk.
37. **Madis Maasing.** The Role of the Bishops in the Livonian Political System (in the First Half of the 16th Century). Tartu, 2016, 318 lk.
38. **Liisi Esse.** Eesti sõdurid Esimeses maailmasõjas: sõjakogemus ja selle sõjajärgne tähendus. Tartu, 2016, 210 lk.