OPEN SOURCE ARCHIVE

Towards open and sustainable digital archives

Liisa Uosukainen (ed.)

MAMK
University of Applied Sciences

OPEN SOURCE ARCHIVE

Towards open and sustainable digital archives

Liisa Uosukainen (ed.)

MIKKELI UNIVERSITY OF APPLIED SCIENCES

MIKKELI 2014

A: RESEARCH REPORTS - TUTKIMUKSIA JA RAPORTTEJA 94

Cover picture: Manu Eloaho
Cover layout: Advertising agency Nitro ID
Layout and printing: Tammerprint Oy
ISBN: 978-951-588-455-8 (nid.)
ISBN: 978-951-588-456-5 (PDF)
ISSN: 1795-9438 (nid.)

julkaisut@xamk.fi

CONTENTS

AUTHORS

Alm Olli, Master of Arts, Information Services and Development Manager, Central Archives for Finnish Business Records

Awre Chris, BSc (Hons) Biochemistry, MSc Information Science, Head of Information Management, University of Hull and Co-chair, Fedora UK & Ireland Fedora User Group

Drake Karl-Magnus, Ph.D. in Geochemistry and Geostatistics at the University of Stockholm, Internal Consultant at the Preservation Department, National Archives of Sweden (NAS)

Juutilainen Matti, D.Sc. (Tech.), Senior Lecturer, Mikkeli University of Applied Sciences

Jääskeläinen Anssi, D.Sc. (Tech.), RDI advisor, Mikkeli University of Applied Sciences

Lampi Mikko, BEng in information technology, RDI advisor, Department of electric engineering and information technology, Mikkeli University of Applied Sciences

Lassila Aki, M.Sc. in information system science, CEO of Disec Oy, previously the Head of Development in the National Library of Finland

Niemi Kalevi, Lic.Soc.Sc., MA, Vicerector, Director of Development, Mikkeli University of Applied Sciences

Palonen Osmo, MA, Managing Director Hyvät Neuvot – Besserwisser Oy / Project Manager Department of Electrical Engineering and Information Technology (retired)

Siitonen Paula, M.Sc. (Comp.Sc.), Head of department of electric engineering and information technology, Mikkeli University of Applied Sciences

Talsi Noora, PhD, Research Director, Digital Archiving and eServices, Mikkeli University of Applied Sciences

Uosukainen Liisa, M.Sc. (Tech.), project manager of Open Source Archive project, Mikkeli University of Applied Sciences

QUESTIONS, SOURCES AND CONTEXTS – DIGITAL ARCHIVING AS A FACILITATOR IN ACQUIRING NEW INFORMATION AND PROVIDING SERVICES

Kalevi Niemi

Discussions on the information society and organizations' information management cannot bypass the challenges of ensuring the availability, reliability and usability of different sources. Digital archiving could be described as crossroads where the new and the old meet in an interesting way, meaning new technology and the traces of times gone by as documents, pictures and sounds.

Historical research has traditionally focused on sources and their value as providing evidence, ie on their critical use. It is, however, obvious that historical sources do not provide researchers with crystal clear truths voluntarily. Sources do not speak of their own will, but because researchers have made questions that make the sources speak. Nevertheless, the critical use of sources does not lose its meaning as the methodological bedrock of historical research, as Kalela (2000) points out. He continues by reminding us that this means further emphasis on the dependence between the purpose and the contents of the source, ie whether the source is really authentic and truly relates to the topic under study.

Also the expectations relating to open data and making public materials available to everyone can be seen in the same way. Documents that first seem silent and passive could be sources for new innovations and services when seen through new eyes. In general, it is important to understand in all archiving work that the availability, reliability and usability of different sources do not serve the researchers only.

But, what distinguishes a document from any piece of text? The underlying criterion is the information about the context where the document was created, with the details of who created the document and when. Especially the era of the internet requires that thorough attention is paid to evaluating the reliability of the information and to ensuring the details of its origin. Information technology isn't therefore replacing the work of archivists, although archives as concrete rooms and buildings will become less important. Document management and archives will still be relevant in ensuring that the information is correct, usable and well preserved no matter whether the operational environment is digital or based on paper. (Lybeck et al 2006.)

In Finland the documents of public administration have already been produced in electronic form for a quarter of a century, but the major part of archiving still takes place in paper form. It is absurd and financially inefficient that materials with originally electronic form are printed on paper to be stored in an archive, from where it is again taken to be digitized into electronic form. This is what the National Audit Office of Finland reports in its recent report. The need for change has been recognized, but there are not many operators to carry it out.

The OSA (Open Source Archive) project of Mikkeli University of Applied Sciences (Mamk), carried out between 2012 and 2014, has created an open source service model for the reception, management and distribution of archive material. The purpose of this present publication is both to introduce the results of this project and to raise new questions and development targets. Mamk already has traditions in developing digital archiving, and as Osmo Palonen does in his article Winning model built at Mamk: tailored services from a common platform, the history of this work can be described through three different periods. Palonen, who was involved in the development right from the beginning, also introduces how reliable long-term digital preservation can result in financially profitable business operations.

The OSA project also involved cooperation with memory organizations, such as Elka, Central Archives for Finnish Business Records, whose representative Olli Alm explains in his article titled An open source archiving system for managing business archives how the current digital archives provide user-

friendly tools for making searches that do not require the users to know the organizations whose archives the information is stored in. The publication also includes a seminar paper of Karl-Magnus Drake from the National Archives of Sweden.

In his article Chris Awre from the University of Hull sheds light on why and how open source and the community are success factors in developing digital archives and repositories with Fedora Commons. The article of Mikko Lampi and Aki Lassila discusses the importance of easily finding and utilizing the archived information of memory organizations under the title The harmonization of digital contents: benefits and requirements. Liisa Uosukainen presents an overview to the OSA application development in her article. After this Anssi Jääskeläinen's article Archival UI – an old relic or modern and novel, in turn, considers the changes in the expectations and needs of the archives' customers. This article also pays attention to the viewpoint of the internet generation providing a link to the topic of the last article of this publication titled Open Source in Education. This is where Matti Juutilainen and Paula Siitonen introduce how open source is used in today's IT education at Mamk and how students can be involved in project work and in exploring new tools and ways of teaching and learning.

Explorers do not follow paths that have been carefully designed in advance, but on the other hand, a journey without any kind of itinerary includes the risk of turning into wandering without a destination. Similarly, according to Marc Bloch (2003) the questions made to the sources must be chosen in a flexible way, allowing room for surprises while ensuring that the selection of the questions works like a magnet that captures the dust created during the years and the work of writing and finetuning the documents worth archiving.

Kalela, Jorma 2000. Historiantutkimus ja historia. Hanki ja jää -sarja. Gaudeamus: Helsinki.

Bloch, Marc 2003. Historian puolustus, Edizioni Artemisia, Helsinki (alkuteos Apologie pour l'histoire, ou Métier d'historien).

Lybeck, Jari et al 2006. Arkistot yhteiskunnan toimiva muisti. Asiakirjahallinnon ja arkistotoimen oppikirja. Helsinki: Arkistolaitos.

A WINNING MODEL BUILT AT MAMK: TAILORED SERVICES FROM A COMMON PLATFORM

Osmo Palonen

There is a paradox built in digital archiving. This was the result of the evaluation I made almost immediately, when I started the last of my three careers as a policy maker and ideologist for Mamk's digital archiving. Earlier I had worked as a journalist from 1970 to 1989 and as an industrial automation sales and project manager between 1989 and 2003. Since starting at Mamk (Mikkeli University of Applied Sciences) I have seen that there are two different digital archiving processes and missions inside this field: The first one is to ensure that digital information remains original, including the documentation of all the events involved in the preservation process, even when this process has been going on, and will be so, for centuries.

The other one is to make digital information available as widely as possible in a format that is the best for the users today. That was the reason why there were two different subjects in the OSA (Open Source Archive) project (OSA project, 2014) plan made at Mamk in 2011. This article concentrates more on the first one, the preservation, as it was my primary interest most of the time when working in Mikkeli. Retirement in June 2014 has not changed this, and the present article is a part of my continuing interest in this topic.

Due to my background as a historian, this article includes more narration than exact numbers compared to the texts by natural scientists that are often involved in the projects of digital archiving. Also at Mamk digital archiving has been developed in relation to its IT education and activities. In fact, I have often questioned the practise that information technology is classified among

natural sciences as a field of education. Perhaps this classification only applies to the hard core of information technology.

Accordingly, the management of archive material does not have that close connection with history, as is traditionally thought in the field of archiving. If archiving policies are about defining what kind of information is preserved and how it is made available to the interested users, I find this field closer to social sciences. The role of other sciences related to archiving – history included – is just to be the users of the information preserved. And, the role of IT is just to provide tools for the others. However, understanding history is not a disadvantage for specialists in the field of digital archiving. Understanding the substance is a must in the development work.

This article aims at capturing the first decade of Mamk's digital archiving operations as follows: The next chapter briefly introduces the essence of all the technological choices made and alternatives tested during this decade. Many more service providers, product and version names and technical details could be mentioned and stories behind the decisions told, but my discussion concentrates on the most relevant technological guidelines that we adopted and followed.

After that follows my historical review of developing the digital repository and archiving service centre at Mamk. Historians often try to find periods in the time frame they are studying. In this case the process of a decade – where several projects resulted in a mature and reliable service provider environment – could be described as development in three episodes. These three periods could be titled in the following way:

1. Development I and test production 2003-2007
2. Service provider and developer 2008-2012
3. Development II and expansion 2012-onwards

Consequently, this is how the historical review proceeds after which I introduce some critical elements for successful long-term and permanent preservation and aspects to consider when preparing the digital contents to be handed over to the next generation. To conclude I discuss further ideas and implementations of digital archives and the future.

Technological guidelines

When building Mamk's repository and the services based on it started in 2003 and 2004 there were no examples to follow in Finland. The decision to purchase an IBM Fast T900 disk system and an IBM 3584 tape library for LTO (Linear Tape-Open) tapes with the management system named Tivoli was already made before I started working with Mamk. The results of this pur-

chasing process fit well the policy that started to take shape. When establishing a repository providing services the disk space and tape system are essential investments, and in 2003 – as well as today – there were no other options than to purchase the components from one or more suppliers.

The main principles Mamk's digital archive team adopted and documented later (Stenius & Sirviö, 2007) were the following:

- The contents are held both on disk and on tape.
- The disk storage has to be based on the one-fail-safe architecture, including double controllers.
- The services of all paying customers are provided through clustered servers.
- There has to be an uninterrupted power supply (UPS) system large enough to keep the current stable and to avoid short breaks in the power supply and to give time for the administrators to close the services properly.

The policy decisions were not so easy in case of the server operating systems. We knew very well that in principle we should prefer open source operating systems, but we did not have much experience in building high availability servers using Linux. We also had to make decisions in connection with the medical sector, as the local hospital district, called ESSHP for short, was the first pilot customer for our preservation services. Radiology was widely digitized in Finland at the change of the century, and there were already services available for managing radiological pictures called PACS (Picture Archiving and Communication System) services.

We selected a service supplied by Sectra which operated only on either HP-UX with an Oracle database or on Microsoft Windows Server with a SQL Server database. From my previous job with Honeywell industrial automation I knew that Windows Server 2000 clusters running on the SQL Server database were stable and easy to maintain also via remote connections, if needed. Based on my evaluation also the idea to start using a Unix version which was at the end of the life-cycle seemed a bad idea, especially when it meant being dependent on only one and expensive hardware supplier. We were also just starting the operation and knew that the projects had to be finished; to have more than one platform was not a good idea in that situation. One supporting factor for our decision to favour Microsoft was license fees. Microsoft licences for servers and databases were reasonably priced in 2004 compared to those of HP and Oracle.

We established the first repository by using the principles just introduced and named it as the Digilab storage. Digilab got its first contents with permanent retention periods within the project called ElkaD, which is introduced

in more detail below. This project forced us to consider expanding our basic policy with the following further principles:

- The archival contents must be managed with the products of more than one software/supplier.
- All contents with permanent or defined preservation periods have to be saved onto the archive tapes in two copies.

These policy decisions show that, although every now and then most IT salesmen have been telling that tape is no longer valid media as a storage, we bypassed that sales talk and kept going against the tide. In the archival IT some material can be space intensive, including digitized film, video and sound for example, and the only option is to keep the archival originals on tape. It was, and still is, important to understand the different nature of day-to day IT operations where information stored can be changed even within microseconds and the archival IT where the originals are permanent. Another relevant point was also understood at Mamk: Digital preservation means that the material held in the preservation archive is not available to the users. Only copies are provided. As a result, we further upgraded our tape policy as follows:

- All archival originals are saved on three tapes: one on WORM (Write once read many) tape, and two on LTO tapes.

This was made possible when a new tape library was installed in 2009. We invested in replacing the whole storage system with a one. This new tape library, StorageTek SL3000, was also equipped with double mechanical hardware to ensure the continuity of the service with high availability. The IBM 3584 was left in use, but the supplier's policy of overpricing the service contract for the library made purchasing a new library the only option.

The new replacement disk storage system was EMC Clarion with 50 TB of disc space. The transfer from IBM to EMC went well, but it took a couple of months to complete it fully. After that the EMC system was for the preservation of medical data. The historical archiving services, in turn, were still using the IBM storage that had been earlier expanded up to 70 TB disk space. It's also worth mentioning that there had already been an arrangement with Elka, ie Central Archives for Finnish Business Records, to follow the general guidelines for archiving: One LTO copy – for both the medical and historical data – was stored in Elka's premises that are located about 2.5 kilometres from Mamk.

The next change in the storage system took place in the winter 2014. We first thought that the EMC system could be used longer, the supplier's inability to understand the needs of the customer resulted in a complete change of the system and the supplier. The new IBM Storwize 7000 disk system with 340 TB of disk space also maintains the 100 TB of data earlier stored on tapes and controlled by using the EMC disk extender (HSM) software.

This time also the older and non-IBM systems can be connected to the new IBM solution, if needed. There is also a big improvement in migrating the data to the new system, when the new system can transfer the data during the normal operation and also read the data from the older system, if it was not yet transferred to the new disks. Storwize also provides the option to divide the system network by using SAN for the medical data and iSCSI for the historical archives. That option gives better security for the daily operations in the medical sector and more freedom to develop new archival solutions for historical archives. In addition, with this technology the replication of the whole system to a remote site or to the use of cooperative partners is much easier than earlier.

When the EMC storage system with disc extender was terminated, the StorageTek tape library got more capacity for normal backup and for archival tape production. The supplier's maintenance fee policy also made it possible for us to use an environment that is critical for Mamk's business purposes longer than for the typical five years. Technically tape libraries can be used for 10 to 15 years. As the first IBM disk storage is still up and running after 11 years from the start-up and the manufacturer's support for the controllers only – not needed for the disks – has been reasonable, it looks like the life-cycle of such a system is mainly limited by the consumption of the power and space needed in the server hall, given that there is a number of spare disk units available. As a lesson learned from the previous storage systems, the new IBM was purchased including seven years maintenance contract in advance. This gave us clear understanding of the total cost of ownership for the next seven years.

The first decade of Mamk's digital archiving

When starting to review the history of Mamk's digital archiving, it's worth paying attention to some incidents and developments that played a role in starting the operations. First, it could said that the operations were first developed through projects that were carried out with EU funding (ERDF and ESF) based on the decisions of local governmental authorities. These projects included, for instance, ElkaD (ElkaD project 2006), Kunda (Kunda project, 2007) and Aton (Aton project, 2007). Second, two cooperation organisations appeared, Elka and ESSHP, who had already understood the need for developing digital archiving. Also the city of Mikkeli had developed an interest in becoming a specialist region in the management and archiving of digital information, and Mamk's management of that time supported the ideas of developing such operations.

Mamk had already started to consider, if there would be any need and room for the research and development in digital archiving, digitization and digital content services during the last years of the previous century. Its IT specialists

and administrators started travel the world – Australia, the UK and Norway, for example – to find out what kind of ongoing activities there were and to bring the ideas back to Mikkeli. Mamk's representatives also visited The National Archives of Finland that, however, was not interested in cooperation at that time.

Even though these experiences did not produce that many direct results, a couple of projects started at Mamk with the help of EU funding. The first steps involved the Digilab project where a digitalization laboratory and a digital media environment was established as the first digital archive between 2001 and 2004. Also experts were hired, and in 2003 there were very few who could honestly define themselves as the experts in digital archiving, and I got the job. One of my references was that I tried to purchase a same kind of digital archive for my newspaper as was installed by the magazine Satakunnan Kansa already in the mid-1980s. I didn't get the money for the investment from my managing director, but this showed that I understood why archived information could provide fast and exact information basis for articles, and therefore, also mistakes could be avoided.

The idea to have a laboratory for development and research with no security of continuous funding was one reason why the development projects eventually resulted in services for paying customers. My 14 years of experience in the automation business had some influence on the situation, and the three terabytes of disk space was straightaway doubled in the hope of paying customers. When leaving Mamk the repository had over 300 terabytes of disk space. Supported by the funding from the regional authorities we also made other purchases: software, an LTO tape library, a digitization studio for sound and moving image as well as some special scanners for paper and microfilm digitization. Most of these investments have been in use ever since, some had a shorter life-cycle as technologies have developed and been replaced by new ones.

Development and test production 2003-2007

It could be said that medical data made it all possible, as not long after the start of the Digilab storage in December 2003, the developments with medical data started. The first customer service contract was made in April 2004 with ESSHP, Etelä-Savo Hospital District, for a one-year test period to provide PACS service and storage for radiological pictures as an outsourced ASP (Application Service Provider) solution. It meant that the storage and server environment had to meet the requirements typical of the medical sector. When this test period was extended, the operation model continued and it received wider use, it provided Mamk with a key for developing the archiving services and digital preservation. We started with a development project, but

since then all the extensions and renewals of the storage systems have been made with the help of the medical solutions and to meet their needs.

In addition to working with ESSHP another starting point was in the mammography screening in cooperation with Terveystutkimus Oy. Mammographies are one of the most data intensive radiology exams and the standard IT solutions were not able to provide the service requested. Mamk's advantages as a service provider involved the flexibility of small operations and that it already had gained reputation as a reliable vendor. Through our experiences of migrating historical archives we, for example, already knew that DICOM-based exams and pictures, used widely as de Facto standard in the medical field, would not need mass migration. When DICOM is used, there will be applications and suppliers to convert the DICOMs to a new format, if needed.

However, cooperation with the medical sector was just one part of Mamk's archiving services, and these medical services were separated from the historical archive services into a subsidiary already two years later, when Disec Oy was established in the summer 2006. It took over all customers of the medical sector, and the transfer was complete in the beginning of 2007. The ASP services – called today SaaS – started to grow remarkably and projects that had started out as tests were replaced by standard service contracts with the customers.

Other projects through which Mamk's digital archiving and services were first developed included ElkaD, Kunda and Aton, all funded by ESF, the Finnish Ministry of Education and the regional and local authorities together with Mamk. The ElkaD software developed in the ElkaD project was our first archive application for sound and digital video based on the RDF modelling. The project studied the basis and relating policies for the metadata model and preservation formats. The metadata model was based on the Dublin Core,and national standards and was called ElkaDCore. This model has been extended in the later developments, but the basis is still there. Also preservation formats were defined for sound and video: Broadcast Wave 24 bit 48 kHz and MPEG-2 in 25 and 50 Bit/s in an AVI container. There were only recommendations to follow when these were chosen in 2004, but the decisions still seem to be proper ones.

In the Kunda project PDF/A was adopted when it was available instead of PDF. The purpose of Kunda was to develop an archive service for municipalities in cooperation with Pieksänmaa municipality. For several reasons, not connected to service and software, it did not result in a commercial product. One of the main tasks of the Aton project was to document and evaluate the results of the development so far. Separate documentation was needed, when all development and test production was carried out in projects, and therefore the operations had not been monitored continuously. This documentation was made internally (Stenius & Sirviö, 2007), but there was also an external

Nordic evaluation by the University of Luleå. (Quisbert, 2007). The results of the latter were presented internationally (Quisbert, 2008, pp.184-189), and the evaluation and the seminar article were parts of the doctoral thesis of Dr. Hugo Quisbert (2008b). In addition, the Aton resulted in a plan for the model of Dark Archive (deep archive) to be used in joint operations. This plan was not implemented, but it gave a solid idea for the new, future developments.

Within these projects we also tested digital signatures, smart cards and server verification software for the verification of digital contents. Eventually we solved such challenges by using Profium SIR (later Profium Sence) as an active archival material provider and IBM Tivoli Storage Manager to control the contents on tape. Further developments of that time based on what we had learned: IT specialists could very well write definitions for archive software, but because they did not have archive expertise the products did not work for long-term preservation, only for storages of active use. We learnt that there had to be good knowledge of archives management, not only IT skills, when designing archiving solutions. In addition, we observed that there was an enormous need for education in the field, and we started to arrange education for national needs.

Service provider and developer 2008-2012

During the second period in developing Mamk's digital archiving and services the operations started to be established and they could be expanded. The model of the common infrastructure between the archival services and Disec Oy was proved successful, and I presented this model in the European conference in digital archiving ECA in Geneva. During these years Disec Oy was growing with the turnover of close to EUR 600 000 in 2008, and with the number of 415 000 radiological exams. The turnover amounted to over 1 million in 2012 and the number of new exams close to 600 000 per year. Compared to other organizations providing PACS services, including the public sector, Disec Oy reached the status of the second biggest in Finland.

When moving on to the archiving services there were a couple of new customers, such as three of the four biggest cities, ie Espoo, Tampere and Turku that continued their archiving operations according to practices developed in the Kunda project. Mamk digitized their sound and moving image archives and provided the archive management software and repository as a service called Yksa, our second generation archival data management software. During this period the preservation of moving images was modified based on the model built in the Library of Congress of the United States. Since then it has been MXF/Motion JPEG2000 (lossless).

There was only one bigger project during this period: From 2008 to 2011 the Viva3 project (Viva3 project, 2011) developed the digitization of 3D objects in relation to the famous Astuvansalmi rock paintings as well as applications to present sights in the 3D format. The project studied also the preservation of 3D contents, but there was not a clear solution available. The project also explored moving image archiving and the MJPEG2000 was found to be the primary method.

During these years we made some developments to have better control of the contents than Tivoli had provided. This is why, when the OSA project was planned in 2011, we already had a clear idea about the Dark Archive, we had started to study as early as in 2005 to 2007 within the Aton project. Through Aton we knew that we did not have the resources to develop our own archival management software, and on the other hand, we liked the idea to join a community that had already developed the software. DAITSS (http://daitss.fcla.edu/) was found as a possible candidate. It is preservation software developed by the Florida Center for Library Automation and available through a GPLv3 license. In principle, DAITSS should have been the solution, but as always difficulties would appear when trying to transfer technology from a different part of the world. In this case adopting DAITSS would have required time to learn the technology used to guarantee deeper understanding.

Development II and expansion 2012- onwards

The idea to develop open source solutions to the historical archives was all the time in the consideration. The Open Source Archive (OSA) project was planned in late 2011 and early 2012. The project started in May 2012 with two primary goals: to develop a new Open source platform for archival solutions and test the Dark Archive solutions developed in the community. The results of OSA are reported in this publication by the other writers.

In 2012 the archiving service started operating under the name Darcmedia, and the operations expanded. There were two archives Toimihenkilöarkisto (The Archives of Salaried Employees) and Työväen Arkisto (Labour Archives) that deployed Yksa's third version configuration of the archival content management software. This service got new customers when the City Archives of the Finnish capital Helsinki and the second biggest Espoo started to use it. In these cases there is not much digital contents in the archive, but by using the same Yksa application the users can manage the analogue material stored on a traditional archive shelf and to provide the information for the research and public in the Internet. At the moment Mamk's Darcmedia is a leading software service supplier for Finnish archives. It also maintains unique and large digital contents in the repository where the dark archive is still based on tapes controlled by IBM Tivoli.

As this historical review has reached the present day, it's time to move on to consider further what the first decade of operations has taught us, and what the future looks like. The next sections, therefore, introduce certain topics that I find key elements of permanent preservation. These include the number of copies and standards, for example. In addition, the last part of the article discusses points to remember and to pay attention to when developing digital archiving and related services.

Elements of permanent preservation

The ideas for preserving digital contents has changed during the 11 years of digital archiving I was working with Mamk. The biggest change might be in the solutions to maintain the integrity and authenticity of data. The first wave in terms of these involved using digital signatures, and lot of hopes that the copies now only on tape could be written onto other media that could not be altered, CD or DVD for example. However, when technology has evolved, and the costs of tapes have decreased, these are no longer challenges in the same way. With LTO-4 tapes' capacity of 800 gigabytes, not to mention LTO-6 with 2.5 terabytes per tape, it's no longer an issue to store several copies on different tapes in two different sites. Taking many copies also helped to discard digital signatures as checking the integrity and authenticity could be achieved through simple checksums. For example, the Tivoli software used by Mamk also checked all the files on tape when the tape was taken into a drive.

Moreover, new communication technologies and new storage features also provide security. If the archive needs to be available during major communication interruptions, too, a complete system or its part can be replicated to a cooperation partner located far enough. However the customers are not often willing to pay the extra cost created by this on line replication. Instead in a Dark Archive this can be done with less complexity, when the time for providing new copies is not relevant. The data between these two independent operations can be transferred by a separate transfer process even by sending the information in an off line package.

Using standards is a further tool to ensure the preservation of contents. The OAIS reference model has been widely used in the archives, although the model is so general that nearly everyone can agree on it. In the case of Mamk the main principles of the model are followed except for the idea to create a distribution copy of the archiving package for every request. Instead, the distribution copies are continuously available for the users and updated based on the change of technology. In the original environment (NASA) where OAIS was designed the data was completely different from the one stored in typical archives.

In 2004-2005 when the first decisions concerning the preservation formats were made at Mamk, there were not many examples to follow. PDF was seen as unchangeable format for standard documents, scanned or converted from the digital originals. During the conference trips we became more familiar with the PDF/A, and it was deployed as soon as the tools were available. From our point of view TIFF was never a preferred format, while when published, it is still under the control of Adobe, and the uncompressed files are large. The problem with lossless compressed JPEG 2000 has been the slow increase in its popularity.

Currently this kind of standard decisions are easier than earlier, as organisations can just select the ones they need and prefer to use from the national portfolio of preservation and ingest formats. I was a member of a working group as Mamk's representative within the KDK (National Digital Library) project and was able to have some influence on these formats. Regardless of the word "national", this project did not invent a wheel: The basis was taken from Canada, and a big part of the standards have been developed in the USA. In standardization it is important to create the international contacts, because there is no need for national standards, when the problems are equivalent.

When turning to aspects that relate to organisations and resources, Mamk's archiving service was evaluated in the end of 2007, as introduced above. This work applied parts from two evaluation tools called Trac and Drambora. It is important to record and evaluate the processes the same way as in quality work against the ISO standards and review the processes regularly. Another important evaluation target is how organizations support repository operations and how committed the owners and administrators are to provide resources. Still, it the most important for operations such as Disec Oy and Darcmedia archiving services to be able to finance their own operations: to have customers and partners who value the service worth the fee they pay for it.

Even though trusted repositories are also evaluated in terms enough resources, too big personnel costs will be more precarious. A much more financially stable situation can be created by common infrastructure like in Mamk's case where the storage is shared and the maintenance is based on cooperation. It is important that the personnel is trained to manage a wider selection of tasks. Creating too expensive organizations is often typical in the public sector, and in the end, the costs will be paid by the customers and users.

When considering the two sides of the digital archiving – the daily archiving management software and the long-term preservation – it is much easier to get the former to cover the costs and make small profit than to make the latter self-supporting as a separate line of business. To keep the contents in a dark archive is not very expensive, if there are not many – if any – dedicated persons for this operation, and the software is primarily bases on open source.

Based on our latest experience a simple data migration between the media, for example, will not cause major costs when changing the storage. However, the price of the full migration of the contents from one format to another can create unknown costs, when the processes have to be carefully designed and the results tested. That is why it is important to find formats that can last longer than the traditional office formats do. For instance, the promise for PDF/A is 30 years. In that case migration should be done only once in an employee's career in the best case. And again in migration, it is wise to find out what the others are doings and to cooperate. The same processes can be used in various repositories and the inconvenience of proper testing can be shared.

When I started to work for Mamk eleven years ago, there were many archivists who said that digital archiving would not be possible. Today this kind of comments have disappeared, but that does not mean that the public administration and civil society have completely adopted the idea of digital contents as permanent compared to contents printed or exposed on paper. For me digital archiving has always been an opportunity. Its first benefit bases on the format and archiving services. Instead of travelling and sitting in the reading rooms the interested scholars or ordinary enthusiasts can use the archived material in their own office or at home, provided that the material can be found, of course.

The other advantage is more professional: Instead of building and maintaining huge special buildings where papers can be preserved for centuries and longer with the help of air conditioning, electricity, heating and cooling, the material can be preserved in not so huge server rooms ensuring that the contents will be preserved by distributing the information nationally and internationally. During my first international conference I met a colleague from Australia, and we agreed on the idea that sharing our material in down under Australia and on the most stable bedrock of Europe, in Finland, would be a good idea. When also incorporating the Canadian Shield (http://en.wikipedia.org/wiki/Canadian_Shield), the international network of preservation could be ready.

The sceptics against digital archiving can be very well understood due to unsuccessful stories such as Vapa, Valda, Kanta and ERA (electronic records archive in the USA) and so on. If it took ten years and EUR 400 million to build a central archive for the national medical records with most of the contents text-based information within a XML structure, how could ordinary citizens think about affording a digital archive for their own records, photographs, digital videos, diplomas and certificates?

Should the development of digital archiving always be like that? And the answer is "No". The option – when compared to these miserable stories – is to be clever and flexible, to use international norms and to avoid major software developers and complicated processes. Don't do anything alone, because there

are always others who have the same problem. Working together it is possible to reduce the costs and to support each other. The national archives of the USA learned from their ERA project not to undertake that kind of huge overall project where the leadership of the project is on the supplier's side, while the customer cannot manage the mammoth. It is better to be agile, outsource the development work to those who really understand exactly that part: eat the elephant piece by piece and have the responsibility in your own hands. (Phillips 2012.) Also another major content holder in the USA, the Library of Congress, shares the same kind of ideas (Brunton & Zwaard, 2013).

Keeping it simple

Going again back to the start of my archiving career, one of my mentors was Matti Lakio, the director of Elka at that time. Matti often argued that "it is not the records the archive is preserving – it is the information the users are looking for". When this is the case, do we really need the information divided in printed pages and preserved as PDF/A files or could it be enough to have the information including the metadata? By using metadata the connection between the information and the record that existed can be confirmed. I know how the text looks like when printed in Arial or in Times New Roman, and I also know that there are contracts and declarations of independence that have to be preserved in their original form. Still, I'm not so sure, if all the texts I have written would be needed in fuller forms than in sequence of characters. A few weeks ago I made a calculation to find out how much text-based information on the average A4 pages could be written onto two 4 terabyte disks. I had just purchased the disks for my home server and ended up with characters worth of 12.8 billion A4s. The economical solution to save characters instead of pictures of A4s has another benefit: the compatibility with the sacred Open Data.

My calculation example covered simple text only. Of course, I know that an art books need to be preserved digitally, so that the look and the feel of the book are preserved. We also like to hear from sound recordings how the city council of Tampere argued about various developments. Even when the automatic transcription would be better for searching purposes, it never brings the tone and real conversation live again. However, at the moment most of the archived material is simple text.

I just read that preserving 200 gigabytes in the VAPA archive of the National Archives of Finland costs EUR 400 000 per year (Ollakka & al, 2014). Mamk proposed the National Archives cooperation, when this ingest and service system VAPA was planned. We knew for sure that the small amounts of data available in the first years of this kind of system could have been stored much cheaply in our existing repository and managed by software using the SaaS

principle. We never thought to gain EUR 400 000 per year, more likely 40 000. It could be further commented that VAPA is an archival service application for the ingest and distribution of born digital material with permanent preservation only. In other words, even the starting point was wrong, when no material with a retention period of 50 years for example, was not accepted in the system.

Digital preservation is a complicated task, if considered so. Separating the preservation and archive services is one way to keep it simple. The DAITSS system of the Florida universities is one example of building a system just for the preservation. The Mamk model that uses just tapes based on published technology and commercial management software of a well-known vendor is another example. The target we set for the OSA project was to start forming a consortium for building an open source alternative for commercial tape management software. This target was not reached. However, there are at least two open source backup software for tapes, but the tape file system LTFS published only a few years ago still seems the most promising. And, I would not be surprised if there would be already plans to use Fedora (Flexible Extensible Digital Object Repository Architecture) Commons (as the front end of LTFS. If not, Mamk needs to put some effort in it. One of the biggest threats in using proprietary software like Tivoli is the imperative role of the system database in identifying where on tape the contents are.

Even though this article has concentrated on the preservation issues, some comments must be made on the other side of digital archiving coin, too. The idea of OSA and its predecessors has been that the customers would be using common resources, servers and instances. By paying extra customers have got an instance of their own, but Mamk's policy has been to support a common hardware and software platform, if the capacity will not be a problem. The other policy has been to build applications which are used by both the archivists and the customers of the archives and which provide access to both the digital objects and the papers and photographs stored in the traditional archive vaults. Furthermore, different archives, archivists and customers using either public contents or contents with limited usage have been separated by the user rights and roles and the logical structure of the archive. The access information is stored in the metadata of the contents. There is no need to build different applications for different users or contents. One size fits all when the outlook can be tailored for every customer and need.

One reason why the OSA project was established was the known risk of variable costs and disappearing elements of commercial software. In the OSA project this goal will be reached by using proven components and the agile development method. The earlier systems developed in Mamk have used some portions of commercial software, although Yksa 3 is just running on a commercial operating system and uses commercial database. During the past

decade the policy to minimise the platform costs between all customers by using a common platform and common storage resources for many different types of data has been very successful. It has made it possible to provide even small archives with professional archive application services by using the SaaS model with a reasonable price tag. The only major change I can see in the future is expanding the customer base in Mikkeli and in finding new users for this model abroad.

The previous sections have described with an example how to establish digital preservation services in a research and development environment. Mamk and its owners have given exceptional support and had interest in making this process possible. The challenge for the current organization and the next generation is to keep this development and service alive. The stakeholders' continuous interest and the operations of the spin-off company Disec Oy are still the key factors in ensuring the progress in the development.

This is where I rely on Mamk's directors and employees and the representatives of the city of Mikkeli. Compared to the activities carried out elsewhere with big money and with not so many results, we can be proud of our development in providing these services. A further important development in this field within Mamk involves the educational programmes that concentrate on digital archiving and content management. Their history was not included in this article, and should be recorded by others. To conclude the long-term preservation of digital information is a task where every generation can just do their best in saving the digital marks of life and the civil society to those interested in their history or in utilizing the old material within new developments in the future. This can be called archiving.

References

Aton project 2007. Aton – arkistoinnin tekninen osaaminen. WWW document. http://www.mamk.fi/aton. Referred 15.10.2014.

Brunton, David & Zwaard, Kate 2013. Repository Development at the Library of Congress. Archiving 2012 Final Program and Proceedings. (Washington DC, USA, April 2-5, 2013).

ElkaD project 2006. ElkaD – arkiston erityisaineiston digitointi, tallennus ja tietoverkkopalvelu. WWW document. http://www.mamk.fi/elkad. Referred 15.10.2014.

Fedora Commons 2014. Fedora. WWW document. http://www.fedora-commons.org/about. Referred 23.10.2014.

Kunda project 2007. Kunda. WWW document. http://www.mamk.fi/kunda. Referred 15.10.2014.

Ollakka, Esko, Lahdelma, Pirkko & Vainio, Eveliina 2014. Sähköisen arkistoinnin edistäminen Tuloksellisuustarkastuskertomukset 11/2014, Valtiontalouden tarkastusvirasto. WWW document. http://www.vtv.fi/julkaisu/sahkoisen_arkistoinnin_edistaminen.xhtml. Referred 23.10.2014.

OSA project 2014. OSA – avoimen lähdekoodin arkisto. WWW document. http://www.mamk.fi/osa. Referred 16.10.2014.

Phillips, Megan 2012. Lessons Learned from NARA's Electronic Records Archives Project. Archiving 2012 Final Program and Proceedings. (Copenhagen, Denmark, June 12-15, 2012).

Quisbert, Hugo 2007. Evaluation of the Mikkeli as Trusted Repository. Mikkeli University of Applied Sciences. WWW document. http://www.mamk.fi/instancedata/prime_product_julkaisu/mamk/embeds/mamkwwwstructure/17872_1805-1271-The_Audit_of_Mikkeli_as_Trusted_Repository_final.pdf. Referred 17.10.2014.

Quisbert, Hugo 2008a. Evaluation of a Digital Repository. Archiving 2008 Final Program and Proceedings. (Bern, Switzerland, June 24-27,2008).

Quisbert, Hugo 2008b. On Long-term Digital Preservation Information Systems – A Framework and Characteristics for Development. Luleå University of Technology. WWW document. http://epubl.ltu.se/1402-1544/2008/77/LTU-DT-0877-SE.pdf. Referred 17.10.2014.

Stenius, Mårten & Sirviö, Heikki 2007. Digital Archive System. Data security arrangements - Systems and policies. Mikkeli University of Applied Sciences. WWW document. http://www.mamk.fi/instancedata/prime_product_julkaisu/mamk/embeds/mamkwwwstructure/17871_1805-1271-DigitalArchiveSystem_20071214_op.pdf. Referred 17.10.2014.

Viva3 project 2011. Viva3. WWW document. http://www.mamk.fi/viva3. Referred 16.10.2014.

TWO APPROACHES TO DIGITAL LONG-TERM STORAGE AND PRESERVATION AT THE NATIONAL ARCHIVES OF SWEDEN (NAS) AND THE CHALLENGE OF FILE FORMAT STANDARDS

Karl-Magnus Drake

1. Introduction

The National Archives of Sweden (NAS) have received transfers of digital records from various state and regional agencies ever since the 1970s. In the 1980s we didn't have any of our own equipment, we used service firms and the transfer control was made from data printouts and paper metadata. However, from the 1990s onwards we have been able to use more resources to develop our electronic archives management and to set up digital repositories to store and preserve digital deliveries from state and regional agencies as well as digitized archival collections of own production.

This paper will first present some digital production lines used when NAS is or has been running various scanning projects during the last two decades. These digitization activities within NAS have been generating a vast amount

of digital objects and metadata now stored and preserved in our digital repositories. NAS' current OAIS compliant long-term digital storage and preservation system and its components will then be outlined as the first approach [1], [2].

Thereafter, a more holistic approach to microfilm-based digital storage and preservation, based on the results from two European research projects (*Archivator, Milos*) and one Norwegian research project (*Astor*), will be presented. The key component in the presented holistic approach to microfilm-based digital storage and preservation is *Archivator*. It is a technical solution for secure, migration-free long-term storage and preservation of digital data on microfilm. It consists of equipment and processes needed for writing onto and retrieving digital data from microfilm. During the development the need for a truly holistic solution for digital storage and preservation was recognised. Subsequently, EU's Eurostars program and the National Funding Bodies in Norway, Sweden, Germany, Great Britain and Switzerland (not involved any more) and Norwegian Research Council set up the *Milos* and *Astor* project, respectively.

NAS have exclusively been selected as a strategic partner for the *Milos* Consortium, because NAS have already gained good knowledge of and broad experiences in both film recording and film scanning technology when running the internal so called MECOM project [3]. This now finalized project dealt with image quality evaluation of digitized and born-digital still images recorded on black/white and color microfilm as well as the cost analysis for long-term storage and preservation of these images on microfilm.

The outcome is a turn-key solution standing on the three legs of *Archivator*, *Milos* and *Astor* designed specifically for long-term digital storage and preservation requirements. This solution includes all hardware and software components of the *Archivator* system, but also film processing machines, a physical storage solution and storage medium (i.e. persistant and high-density microfilm) to cover a full workflow. NAS have successfully tested the *Archivator* workflow on various digital data objects in an OAIS context.

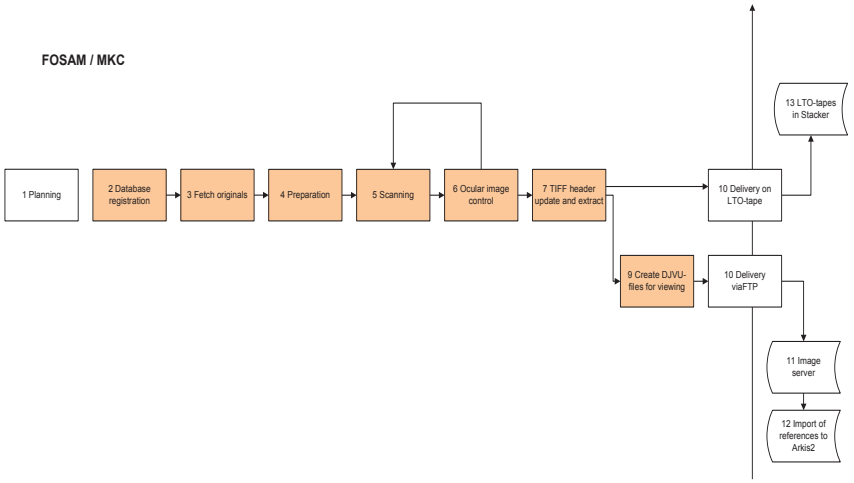
One of the main challenges for NAS' existing as well as the novel microfilm-based *Archivator* storage and preservation system are by no doubt *file format standards,* which will be discussed at the end of this paper.

2. Digital chains for digitizing archival objects

Digital objects generated from five exiting, so called Digital chains are the main digital content currently deposited in NAS' digital repositories.

The five Digital chains at NAS mainly originate from the following digitization projects or activities:

- The FOSAM project: scanning the documents of church records at NAS (younger than 1860), MKC
- The DAF-Access project (scanning highly requested documents at NAS and further processing at MKC or SVAR)
- Ongoing microfilm scanning of various (old) microfilm collections of NAS, SVAR
- The GSU I project started in 2007 (microfilm scanning of Swedish church books older than 1860 by FamilySearch in Salt Lake City, USA and then delivery to SVAR)
- The GSU II project started in 2011 (microfilm scanning of Swedish judicial records by FamilySearch in Salt Lake City, USA and then delivery to SVAR)



PICTURE 1. The Digital chain for scanning church books at MKC within the FOSAM project. After more than ten years of digital production of the documents of church records generated between 1860-1991 the FOSAM project is now finalized. Due to legal personal integrity reasons these digital church books are continuously realeased and web-accessed for the public.

3. Digital repository content and quantities

As can be seen below, the main quantities in NAS' current digital repositories emanate from scanning the paper documents of church records and microfilm reproductions (i.e. the FOSAM and GSU I projects).

- Born-digital files from agencies: 3 TB
- Audio-video files and multimedia: approximately 100 TB
- Digitized volumes (one AIP per volume): 320000
- Digitized images in the TIFF format (mainly from church books): 1400 TB
- Images total (mainly from church books): 121 million
- Images published on the web/Internet: 63 million
- DJVU files (web-presentation files extracted from TIFF files of church books): 18 TB
- Total storage: 2300 TB in a modern HSM-system based on LTO tapes and 140 TB in an older HSM system (obsolete magnetic tapes)

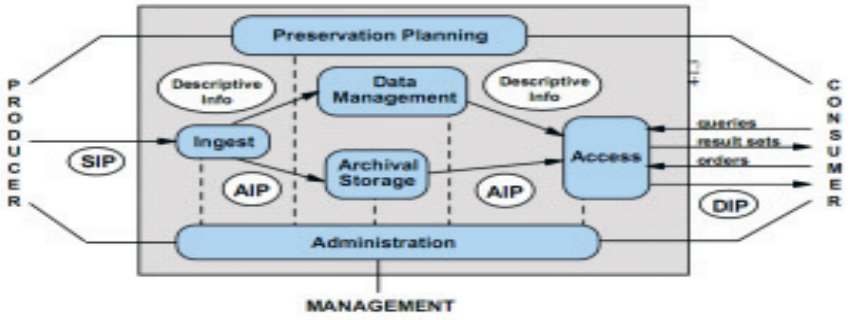
4. Current long-term digital storage and preservation systems

The management of all the above mentioned digital content in NAS' existing digital repositories follows in general an OAIS compliant Long-Term Digital Storage and Preservation (LTDSP) workflow (including the receiving, ingesting, archival storing, data management and accessing processes). These general processes presented below have now been more or less implemented and merged together as various modules in a technical platform called RADAR at NAS. Due to technical evolution and other external factors we expect that further development of RADAR will continue in the future.

Before outlining NAS' current LTDSP system (daily called the electronic archive) the general functions and main requirements of the existing OAIS model and a summary of the formal metadata standards and schemes adopted by NAS are presented.

4.1. The OAIS model – the functions and main requirements

In Picture 2 the basic functions in the OAIS model are delineated [4]. The SIP, AIP and DIP refer to Submission, Archival and Dissemination Information Package normally containing the data to be archived together with all relevant metadata.



PICTURE 2. Basic functions in the OAIS model

The main OAIS requirements for electronic archives:

- Negotiate for and accept appropriate information from information Producers.
- Obtain sufficient control of the information provided to the level needed to ensure Long-Term Digital Preservation.
- Determine, either by itself or in conjunction with other parties, which communities should become the Designated Community and, therefore, should be able to understand the information provided.
- Ensure that the information to be preserved is Independently Understandable to the Designated Community.
- Follow documented policies and procedures which ensure that the information is preserved against all reasonable contingencies, and which enable the information to be disseminated as authenticated copies of the original, or as traceable to the original.

4.2. Adopted archival standards and metadata schemes

General archival standards adopted by NAS

- **ISAD(G)** General International Standard Archival Description and **ISAAR(CPF)** International Standard Archival Authority Record for Corporate Bodies, Persons and Families
  - NAS' Archival information system ARKIS is modelled after these standards.
- **EAD** - Encoded Archival Description [5] and EAC-CPF - Encoded Archival Context - Corporate bodies, Persons, Families [6]
  - Used as exchange formats for archival description information at NAS
  - Supported by several commercial archival information systems
  - Import and export functions in ARKIS
  - Currently a new Swedish EAD and EAC-CPF specification is being developed.

- **OAIS** (see above)
  - Widely adopted in Sweden not only by NAS
  - Several commercial E-Archive system claim to be OAIS compliant

Use of METS at NAS (**METS (2004) -** Metadata Encoding & Transmission Standard [7] - Structure for encoding descriptive, administrative, and structural metadata - DLF/LOC.

- Currently METS is used by NAS to describe a SIP or an AIP.
- Extension schemes in use are ADDML and PREMIS/MIX.
- The current METS profile was developed in May 2010. See profile and description at http://xml.ra.se/METS/
- The METS profile is also used by the National Library of Sweden.
- A similar METS profile is currently being adopted by the National Archives of Norway.

Use of PREMIS at NAS (PREMIS (2005) - Preservation Metadata [8] - A data dictionary and supporting XML scheme for core Preservation metadata needed to support the long-term preservation of digital materials - OCLC/LOC

PREMIS metadata is created by KRAM, the NAS' application for control and ingest of information received from agencies.

- PREMIS and MIX metadata is created for scanned images.
- PREMIS: OBJECT is used for technical metadata and PREMIS: EVENTS for preservation events (for example checksum control).
- PREMIS metadata is embedded in METS by the Swedish Archival storage system ESSArch.
- Metadata derived from PREMIS is also stored in the archival information system ARKIS.

Use of MIX at NAS (MIX (2006) - NISO Metadata for Images in XML [9] – An XML scheme for encoding technical data elements required managing digital image collections - NSI/NISO.

NAS use MIX as metadata format for scanned images in the TIFF format.

- MIX metadata is created by extracting information from the TIFF headers in the scanned images. This is done by the TIFFEdit application developed for NAS.
- MIX-metadata is embedded in METS and PREMIS by the Swedish Archival Storage system ESSArch when AIPs containing scanned images are ingested

Use of ADDML at NAS (ADDML (2001, 2009) - Archival Data Description Markup Language developed by the National Archives of Norway

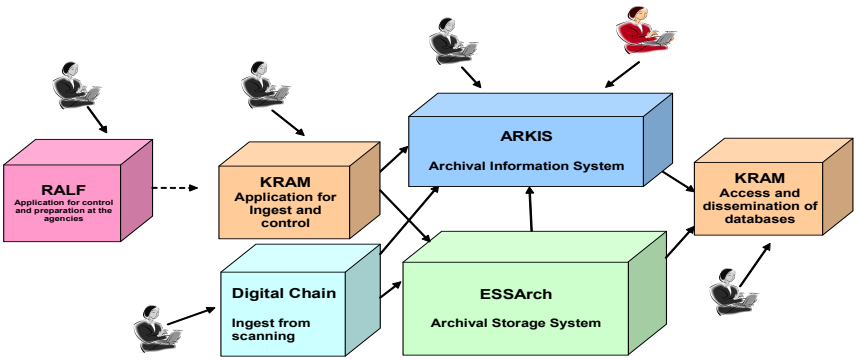
Metadata in the ADDML format [10] is created by the Swedish applications KRAM and RALF for the control and ingest of born-digital information from state agencies

- ADDML metadata is embedded in METS by the Swedish Archival Storage system ESSArch.
- Metadata derived from ADDML is also imported and stored in ARKIS.
- SIPs with METS and embedded ADDML will be delivered to NAS by Statistics Sweden.

4.3. The RADAR platform

The RADAR solution for the electronic archives of NAS mainly consists of the following modules and specifications (Picture 3):

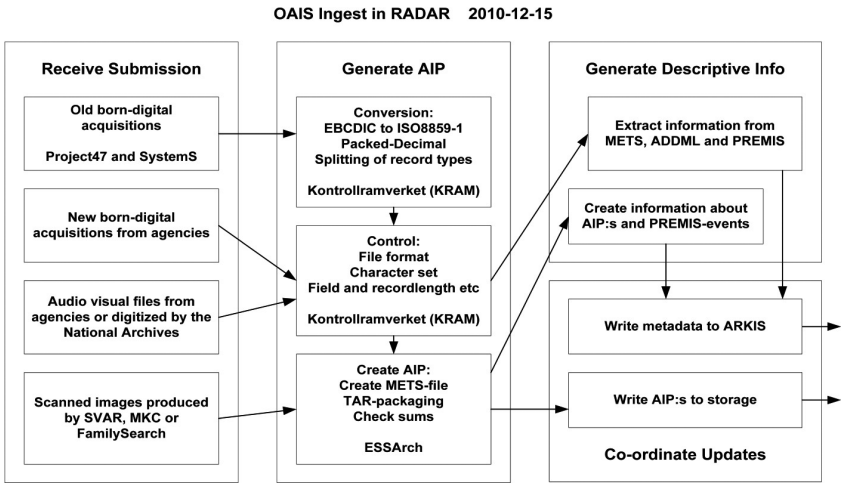
- A general package model (SIP, AIP, DIP) originating from the OAIS model
- Use of general (often XML based) metadata standards such as METS and PREMIS
- Use of specific metadata standards for different types of systems, for example, the ADDML standard for databases
- A self-developed system for transfer control and conversion: KRAM
- A free downloadable self-testing tool of SIPs: RALF
- An open source archival storage system ESSArch: full integration with our general Archival Information System, ARKIS, work processes.



PICTURE 3. The RADAR platform for digital long-time preservation and archival storage at NAS

The main idea behind the RADAR solution is a general and common package structure which has different options concerning the content. The content options are dependent on the source system and are called transfer types, which each have its individual specification.

The Ingest processes in the RADAR system at NAS related to the conceptual OAIS model and the adopted archival standards and metadata schemes are presented in Picture 4.



PICTURE 4. The Ingest processes in the RADAR system at NAS related to the conceptual OAIS model

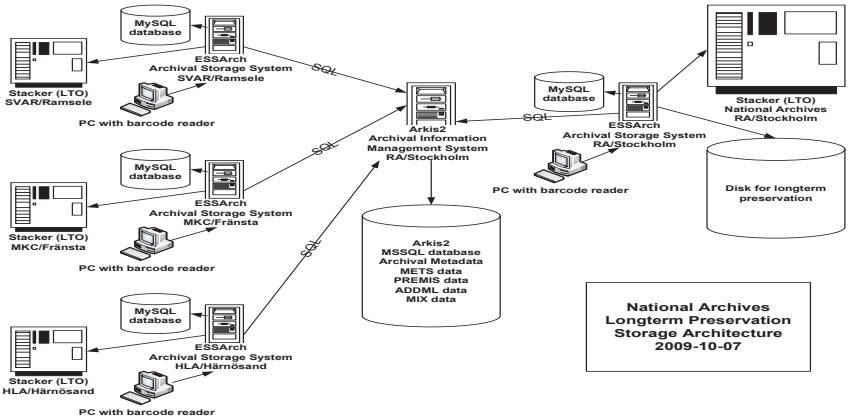
4.4. The main characteristics of NAS' Archival Storage System (ESSArch)

As mentioned above, EESArch is the archival storage module in our long-term digital preservation solution. It manages the packaging, storage and retrieval of the AIPs. Software tools for checksum calculation are used both for the packages and the contained metadata and data files. There is logging of the events taking place. The packages are in the TAR format and may be stored on different media, normally on disk and tape, at present LTO.

- ESSArch is a back end system for digital archival storage.
- Storage and retrieval of AIPs.
- Storage of AIPs in several bitwise identical copies
- AIPs (containg data files and metadata in the METS/PREMIS format) are stored in the TAR format.
- No vendor specific backup format

- Reading and writing checksums for packages and files
- Event log and access control
- Local MySQL database using the PREMIS 2.0. data model
- Automatic updates to the Archival Information system ARKIS
- ESSArch is an open source system based on Linux, Apache, MySQL and Python
- ESSArch (version 2.1.0) is available at SourceForge (http://sourceforge.net/projects/essarch/)
- Used by the National Archives in Sweden and Norway

In Picture 5 below the digital archival storage architecture at NAS (ESSArch) has been outlined.



PICTURE 5. The digital storage architecture at NAS (ESSArch)

5. A holistic approach to microfilm-based digital storage and preservation

5.1. The Archivator Project

An industrial consortium, working together in a project called *Archivator* (now finished), was set up in mid-2009 with the aim of developing a reliable, secure, cost-effective and long-term digital archive solution using photosensitive polyester film (i.e. microfilm) as storage medium [11], [12].

Film in general is a well-established and stable storage medium that remains unchanged for centuries in optimal conservation conditions. The *Archivator* project has resulted in a long-term storage and preservation system - the *Archivator* system is now in beta-testing - including a data recorder, a data scanner and related hardware and software (i.e. all equipment and processes needed to write onto and read data off the storage medium).

In general, long-term digital preservation requires data objects to be stored safely and securely on some sort of storage medium. For obtaining high quality of these data objects as well as automatic retrieval and integrity control in the future, they should be in a binary format. Binary data need to be decoded by decoder software. The decoded bit stream is assembled into files. A viewer for the file format has to be available for future users to fully understand the content.

During the development we realized that the above mentioned *Archivator* project had room for improvements in order to fully cover all requirements related to digital storage and preservation. A more holistic approach is necessary if the true needs of long-term storage and preservation of digital data are to be met. As a result, two new industrial consortiums were set up: the *Milos* and *Astor.*

5.2. The Milos Project

Milos is an ongoing research project funded by the European EUREKA Eurostars program [12]. The consortium is unique in the respect that it represents the majority of players in the European"photosensitive industry" cooperating towards a common goal.

The project objectives are to:

- develop a new microfilm-based storage medium based on nano-sized photosensitive materials for migration-free long-term storage and future retrieval. The storage medium has enhanced storage capacity and is longevity-tested for up to 500 years' durability.
- develop new polymer materials, tested for any impact on the longevity of the storage medium, to be used in the packaging.
- achieve the highest possible speed of data writing and reading on the new medium.
- establish reliable, cost-effective manufacturing methods for the medium and packaging.
- develop the relevant photochemical film processing technology and related equipment that support this new medium.
- prove the longevity and reliability of the new medium.
- prove the concept by integrating the *Archivator* system with storage and preservation systems at existing archives according to the accepted OAIS framework.

5.3. The Astor Project

The ongoing *Astor* project will result in a fully integrated system, pulling a number of other data storage technology projects together to provide one commercially viable solution [12]. The project is exclusively funded by the Norwegian Research Council.

The project objectives are to:

- design a logistical process to secure fast ingest and retrieval of data in an automated robotic data-warehouse, supporting longevity of the integrated system for up to 500 years.
- develop and test polymer materials that support longevity of more than 500 years.
- use these materials to design and manufacture a packaging capable of storing and protecting an optical storage medium (i.e. current relevant microfilms in the market).
- create and test the so called AS4e (Automated Storage Forever) system. All materials in this automatic storage solution are tested for possible effects on the storage medium and primary packaging, ensuring that there will be no negative impact on the projected lifetime of the stored data.

5.4. Key Innovations and Results

Altogether the results from the *Archivator*; *Milos* and *Astor* activities form a unique turn-key solution designed specifically with the requirements of digital storage and preservation in mind. Due to the two latter projects all components will have undergone extensive testing to ensure that longevity of up to 500 years is sustained under normal storage conditions (temperature and humidity requirements). The technology is flexible and open, but has the potential of offering a fully integrated and future-proof solution.

5.5. The trustworthiness of digital repositories

Many national archives and libraries often declare themselves as OAIS compliant to underline the trustworthiness of their digital repositories. However, it is challenging to measure the trustworthiness of digital repositories. Therefore NAS, as a partner in the *Milos* Consortium, has been eager to evaluate if the existing *Archivator* workflow could satisfy the basic requirements in an OAIS compliant digital repository.

In order to establish a fully reliable and OAIS compliant digital repository also the longevity of the storage medium in its packaging is important. Logistics processes to secure fast ingest and retrieval of digital objects should also be considered.

Also an independent body should be responsible for the auditing and certification of such a holistic solution for long-term digital presservation and access. Piql is setting up a certification program for its *Archivator* service providers.

5.6. An Archivator workflow in the OAIS context

Ingest

- Only quality controlled and verified SIPs (Submission Information Packages) are accepted as ingest. The SIP includes digital data objects (e.g. text files, raster and vector still images files, video, sound, data animation or the aggregates of single data objects such as pdf files and web sites –a warc file) and required metadata.
- The ingest function must be able to identify, validate and check the integrity of:
  – Files and file formats embedded into the SIP.
  – Metadata and schemes embedded into the SIP.
- The SIP and additional preservation metadata are used to create the final AIP (Archival Information Packages).

Archival Storage

- Archival storage only accepts quality controlled and verified AIPs.
- The encoded AIP are then written onto the microfilm as square dots by the film recorder together with file format and metadata scheme descriptions. In addition, control data (e.g. film frame layout and indexing data) as well as the source code and technical instructions for decoding the content are written. To design viewers for the archived files in the future file format descriptions and source code are also written to the microfilm. All the documentation for future interpretation of the archived content should be in human-readable form.
- Some digital data objects (e.g. raster and vector still image files and perhaps a small part of a video sequence) could easily be rendered as previews together with the square dots on the microfilm.
- After the microfilm processing all of the content are finally quality controlled.

Data Management

In a migration-based archive solution data management is an expensive never-ending task. This includes monitoring dataflow, maintaining data integrity

and file/metadata migration. Since *Archivator* uses a migration-free medium and stores the file format description together with the content, data management is done once when storing the AIPs on microfilm.

Future access and human understanding

In order to restore (i.e. read and decode) as well as intellectually understand the digital data objects encoded as square dots on the microfilm decades or even centuries ago, future users have to do the following distinct tasks:

- The square dots have to be read/rasterized by a film scanner, digital camera or any appropriate optical reading device.
- A decoding application has to be designed according to the human-readable source code or technical instructions recorded onto the microfilm decades/centuries ago and then implemented.
- The file and metadata format descriptions written onto the microfilm have to be used to design and implement rendering software to interpret the decoded digital objects and metadata.

Applying the Archivator workflow on digital data objects from NAS

Test material

All type of digital data objects (e.g. church books, maps, paper print photographs, technical drawings, illustrated medieval pachements, handwritten letters, modern newspapers in pdf form, seals as well as video and sound) and accompanying metadata in NAS' existing repositories are well represented in the selected test material for evaluating the Archivator concept.

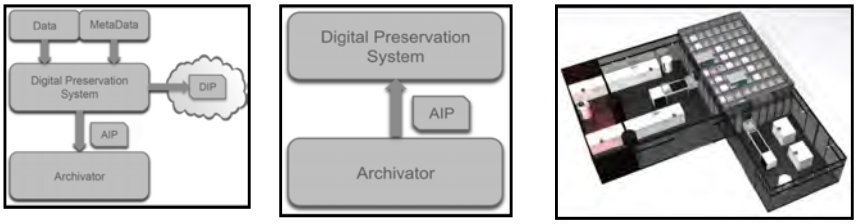
Store (Picture 6a):

- Data and metadata are ingested into an OAIS compliant Digital Preservation System (DPS - e.g Archivematica [13], RODA [14] [] or the existing DPS).
- The DPS validates, checks and normalizes digital data objects before creating a SIP.
- The DPS creates an AIP and DIP (optional).
- Upload the AIP from the DPS
- Prepare the AIP for writing to microfilm
- From the AIP: identify file formats, generate file list and create previews (optional).
- Encode the digital data object to square dots in image frames. The encoding includes Forward Error Correction (FEC) and generates checksums.
- Create the reel structure with a table of contents, human readable section with decoding instructions and fileformat descriptions.
- Write the prepared image frames to microfilm.

- Process the microfilm.
- Put the microfilm in the labeled package. The label includes unique identifiers in both human and machine readable form.
- Verify the microfilm by reading back the frames, decoding the digital data objects and comparing with the original AIP.
- Store the verified microfilm reel in the automated storage system.

Restore (Picture 6b):

- Search metadata in the DPS to access digital data objects.
- Send request from the DPS to Archivator to restore the AIP containing the digital data objects.
- Process the request from the DPS to link AIP with reel in the storage.
- Fetch the reel from the automatic storage system.
- Read the frames from the microfilm.
- Decode the frames to AIP. The FEC is used to ensure data integrity and quality controlled with checksums.
- Download the AIP and store the reel.



a. b. c.

PICTURE 6. a. Ingest and store in the *Archivator* system; b. access and restore in the *Archivator* system and c. a complete, holistic, storage and preservation system

Results from applying the Archivator workflow on digital data objects from NAS

- The digital data objects from NAS were successfully ingested to Archivematica.
- Archivematica generated a SIP as well as an AIP.
- The AIP was uploaded to the Archivator system.
- The AIP with additional Archivator control data was written onto microfilm frames both visually (previews) and digitally (square dots).
- After microfilm processing, the microfilm frames were read back and decoded and the digital dataobjects were reconstructed.
- The reconstructed digital data objects were compared to the originals and found to be binary equal.

In this test we did not use the human readable information on the microfilm to restore the digital data objects. This work has been initiated but not completed when writing this document.

A holistic approach to digital storage and preservation

With the results from the *Archivator*, *Milos* and *Astor* projects, we can now present a holistic approach to digital preservation [9] as outlined in Picture 6c. At its core is the microfilm (e.g. the media) specially designed for storing digital data for centuries. To ensure longevity of the microfilm we have developed new packaging and labeling. The microfilm, packaging and labeling are now being tested together to prove longevity of up to 500 years. On top of this, we have a new high resolution and high speed film writer and film reader (i.e. film scanner). To automate the preservation workflows we have integrated barcode readers to read the labels, a warehouse management system, robotic storage and data management system. In front of this we connect to existing digital storage and preservation systems.

6. The Challenge of File Format Standards

6.1. Background

Besides large-scale disasters (floods, fires, earthquakes and acts of war), media faults (physical degradation and bit rot) and media/hardware obsolescence memory institutions worldwide are facing a number of "software" challenges when ensuring future long-term accessibility, readability and human understanding of the context and content of digital objects which are today stored and preserved for various legal; scientific and other reasons.

The two main components of these digital objects are 1. metadata in various schemas (see above) 2. the data content (which could also contain metadata to a certain degree) encoded and sometimes compressed according to a variety of file format standards and compression schemas mainly depending on the nature and usage of the content.

A file format standard is in practise a well-established and authorized document specifying in technical terms the details necessary to construct a valid file of a particular type and to develop software applications that can decode and render such files. The actual specifications may vary considerably in length, from well under 100 pages to well over 1000, depending on the complexity of the format.

A format specification should indicate the sequence of the different units/subunits in the file framework, and whether a particular unit/subunit within the bitsteam should be interpreted as alfanumeric ASCII characters, binary machine instructions, color data or something else.

6.2. File format obsolescence

Alike media/hardware obsolescence metadata schemas and file format specifications can also become obsolescence sooner or later for a number of reasons:

- proprietary, closed/open format specification (i.e. not ISO-standarized)
- limited number of applications (e.g. editors, viewers, web browsers)
- few application developers/vendors
- applications supporting the format no longer exist
- the format is withdrawn due to copyright dispute; commercial or other reasons
- limited adoption
- not compatible with e.g. current hardware/software/OS platforms
- no/limited backward compatibility
- limited (production) metadata support
- limited range of functionality
- very complex functionality (e.g. only editors created the file could read ditto)
- limited upgrade cycle
- limited technical documentation

6.3. The file format implementation dilemma

The file format dilemma starts when memory institutions in charge of the long-term storage; preservation and accessibility of digital objects and accompanying metadata are implementing even well standardized file formats (e.g. ISO file formats) in their OAIS-modelled digital depositories as SIPs/AIPs (i.e. Submission/Archival Information Packages). As the applications implementing the file format selected by the information owners (i.e. the memory institutions) or managed by internal/external service providers are not in control neither by the information owners or the service providers the file format compliance cannot be fully guaranteed. Tests to check compliance with file format standard specifications could indeed be performed by means of various file format validating tools. However, these validators are not fully reliable as different tools could end up in different results. This poses real problems

in long-term preservation and the future access of digital objects physically stored on *all* type of data carriers (LTO tapes, optical discs as well as the microfilm-based media used in the *Archivator* system).

If the outcomes of a compliance test do not provide positive results the processed files have to be returned to the information owner/service provider for corrections and the SIP/AIP generation process has to start again, and hence, the costs easily rise out of control. In the worst case such uncontrolled SIP/AIP generation process could jeopardise the whole long-term storage and preservation workflow, not to mention the risk for complete loss of readability and intelligibility of decoded digital objects in the future.

Even if the *Archivator* system records all required technical specifications of file format standards concerning stored digital objects in human readable form on the microfilm-based media, it does not help when trying to access these digital objects in the future, if the digital objects are not fully produced according to the referred file format standards.

People in charge of the technical implementation and the daily management of the *Archivator* system always have to consider that in reality there could be a fundamental difference between a specification of a standard and the implementation of that standard in an application. Moreover, there is always uncertainty concerning the extent in which a specific standard is implemented in a specific application (e.g. a pdf file viewer).

To get full credibility concerning future accessability; readability and human understanding of digital objects stored today on NAS' LTO tapes as well as on the microfilm-based media in the *Archivator* system the service providers have to guarantee that what is produced according to a file format standard is also in full compliance with the standard specification in question!

There is no easy solution to address the file format implementation dilemma for NAS' "electronic archive" or the *Archivator* system but you have to primarily consider the above list of file format obsolescence and well-established sustainability factors outlined in Table 1 below exemplified by the three still image JPEG2000 file format standards [15]. Moreover, you always have to apply file format validation and conformance tools to your digital data objects to be ingested in a long-term storage and preservation system.

TABLE 1. Evaluation of the fulfilment of the sustainability of three still image JPEG2000 file formats standards (Scoring: 1=not so good; 2=good; 3=very good)

Sustainability factors:	JP2	JPX	JPM
1. An ISO standard file format	Yes	Yes	Yes
2. Disclosure			
- Copyright-free, available, complete technical specifications and frequent updates	Yes	Yes	Yes
- Tools to perform a reference implementation	Yes	No	No
- Tools for compliance tests	Yes	No	No
3. Licensing and patent claims	No (?)	Yes	Yes
4. Adoption			
- Adopted by many users and deployed in many disciplines	2	I	I
- Implemented in many applications	2	I	I
- Can be implemented on many hw/sw platforms	3	3	3
- Many development tools available	2	I	I
- Format identification, validation and characterization tools available	Yes	Yes	No
5. Self-documentation			
- Support for various embedded metadata	3	3	3
- Support for most common color space and embedded ICC color profiles (will be changed soon)	I	3	3
6. Transparency			
- "Simple" coding	2	I	I
- "Simple" compression	2	I	I
- Option for technical "lossless"compression	2	I	I
- Effective usage of network bandwidth -	3	3	3
- Robustness when networking	3	3	3
- Protection for data corruption	3	3	3
- Many options concerning image rendering	3	3	3
- Option for copyright management (Note: violating transparency!)	3	3	3
7. External dependencies			
- Dependent on specific hardware	No	No	No
- Dependent on specific development tools	No	No	No
- Dependent on specific OS	No	No	No
8. Technical protection considerations			
- Dependent on specific implementations	No	No	No
- Dependent on complicated coding and compression Schemas	Yes	Yes	Yes

7. Conclusions

A comparison of the main strengths and weaknesses of the *Archivator* storage and preservation system and the NAS' "electronic archive" (*NAS-EA*):

- *Archivator* is a migration-free (or at least there is less migration) off-line solution, and hence inexpensive from a long-term perspective. *NAS-EA* is based on repetitive migrations with negative implications on long-term cost.
- *Archivator* is based on microfilm-based storage media with a proven longevity of up to 500 years compared to the current LTO tape solution of *NAS-EA.* The estimated longevity of (magnetic) LTO tapes is only at the most 30 years (hence the need for repetitive migration).
- *Archivator* is based on non-magnetic (optical) storage media (i.e. microfilm) compared to the magnetic-based media solution of *NAS-EA* (i.e. LTO tapes) which could be exposed to external threats such as electricity shortage, electromagnetic pulses (EMP), etc.
- *Archivator* stores and preserves digital data objects on a media (i.e. microfilm) which cannot be edited, changed or altered. Once the data is written they are like carved in stone (true WORM) in contrary to the rewriteable LTO tape media of *NAS-EA*.
- *Arcivator* data is written in digital form onto microfilm media, just like it is done with any other digital storage media like magnetic media (such as LTO tapes) currently used by *NAS-EA* . This means a unique permanence and zero quality loss at optimal storage conditions in both cases.
- *Archivator* is the only solution that allows preserving a combination of digital data, human-readable text and visual information/pictures on the same medium. Such a hybrid solution reduces the need for using multiple technologies and processes, and is thus increasing the efficiency as well as the reliability of the data storage and preservation.
- *Archivator* allows storing metadata (i.e. descriptive information about the contents) directly onto the storage medium as well as online. Integrated into an IT environment, this means that you can easily search and find the data you need on one and the same physical media (i.e. a microfilm reel) as the data is stored (no broken links between data and required metadata). This is not so easily achieved by the LTO tape solution of *NAS-EA.*
- *Archivator*, as well as *NAS-EA*, addresses the file format standards implementation dilemma.
- *Archivator* (currently) stores archival content at a higher initial cost per GB than the LTO tape solution of *NAS-EA*.
- *Archivator* always stores archival content on an offline storage medium which is more challenging than the LTO tape solution of *NAS-EA* when accessing information.
- *Archivator* has a relative long work flow requiring chemical processing etc., which is not the case in the LTO tape solution of *NAS-EA*.

8. Acknowledgements

I want to thank especially my colleagues Mats Berggren, Magnus Geber, Karin Bredenberg and the staff at the NAS-EA ("the electronic archive") unit, all at NAS, for contributing to the content in chapter 2-4. Finally I also thank the main partner of the current MiLoS project, Piql A/S (former Cinevation A/S), for contributing to the content in chapter 5 and 7.

References

[1] Berggren, M., 2011, Digital long-time preservation and the use of standards at the Swedish National Archives, Expert Meeting, Digital Preservation at the European Commision, Brussels

[2] Geber, M., 2012, Transfers and preservation of e-archives at the National Archives of Sweden, ICA 2012 Congress, Brisbane, Australia

[3] Drake, K-M., 2008, Methods of long-term storage of *E*lectronic Archives on *COM* (MECOM), Internal Technical Report in Swedish, The National Archives of Sweden (NAS).

[4] ISO (2003) Open Archival Information System (OAIS), ISO 14721:2003. (2012-03-29). http://public.ccsds.org/publications/archive/650x0b1.pdf

[5] EAD (Encoded Archival Description) http://www.loc.gov/ead/

[6] EAC-CPF (Encoded Archival Context - Corporate bodies, Persons, Families) http://eac.staatsbibliothek-berlin.de/

[7] Library of Congress (2010) Metadata Encoding & Transmission Standard (METS). http://www.loc.gov/standards/mets/

[8] Library of Congress (2012) Preservation Metadata: Implementation Strategies (PREMIS). http://www.loc.gov/standards/premis/

[9] MIX (2006) - NISO Metadata for Images in XML. http://www.loc.gov/standards/mix/

[10] Archival Data Description Markup Language (ADDML) http://www.arkivverket.no/arkivverket/Arkivbevaring/Elektronisk-arkivmateriale/Standarder/ADDML

[11] Plata, O., Bjerkestrand, R., 2012, The ARCHIVATOR - A solution for Long-Term Archiving of Digital Information, Proc. Archiving 2012, IS&T, pg. 70-74

[12] Brudeli, B., Drake, K-M., 2014, A Holistic Approach to Digital Preservation, Proc. Archiving 2014, IS&T, pg. 79-83

[13] Archivematica https://www.archivematica.org/

[14] RODA http://www.roda-community.org/

[15] Drake, K-M., 2011, JPEG2000-standardens uppfyllelse av kriterier för stillbildsformat för långtidslagring av Digidaily-material. Internal Technical Report in Swedish, The National Archives of Sweden (NAS).

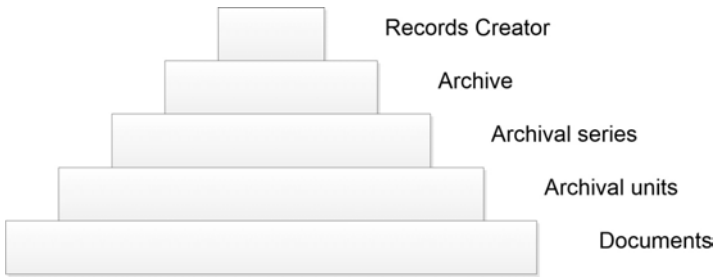
An Open Source Archiving system for managing business archives

Olli Alm

Central Archives for Finnish Business Records – abbreviated as Elka – is a national depository for business documents, and the most important research archive for business history in Finland. Elka's customers include companies that need storage, digitisation and information services, but also individual researchers and other parties needing information. Today, most information in companies is managed digitally, and therefore Elka, too, must be able to provide reliable and high-quality digital services to its customers. Our new open source archiving system enables us to offer even more innovative, open and, most of all, user-friendly services to all our customers.

The traditional way of managing archives

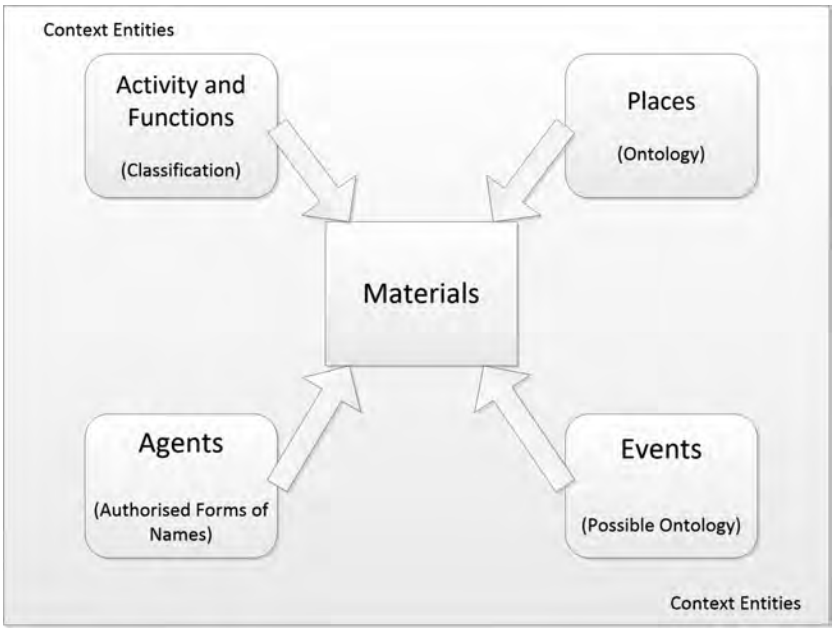
Traditionally, the Finnish paper archives have been managed by creating hierarchical descriptions and lists. As Picture 1 below shows, the description has consisted of levels known as the records creator level, archival level, series level and archiving unit level. Such an archive is the result of the activities of a single records creator, and contains documents created and received by the records creator (Arkistolaitos 1997, 11). Therefore, the relationship between the archive and the records creator has been very close – one archive has only contained documents from a single records creator (Arkistolaitos 2006, 16–17). On lower levels the archive is divided into series and the series into units. Although logical, this kind of pyramid model is inflexible, and poorly suited to modern information management (Alm & Strömberg 2013, 9).



PICTURE 1. Traditional pyramid model for archival description

The model of contextual description

From the viewpoint of an archiving specialist, the most significant and interesting aspect in the new open source archiving application is the revised description model. OSA, the Open Source Archive system, can implement the model of contextual description developed in the Capture project in 2011–2012. This model abandons the traditional hierarchical pyramid. It was first published by Alm and Lampi (2014), and is introduced in Picture 2. The description is implemented through independent entities linked to each other. The main entity – the entire object of description – is of course the archive material itself. The descriptions of the materials focus only on describing the documents and their content by using, for example, dates, subjects and descriptions. The traditional records creator is replaced by a separate entity called agent. This agent can be any natural person or organisation that has influenced the production, collection, storage or the use of the materials.



PICTURE 2. Model of the contextual description

The materials are divided into smaller subunits, such as groups and archive units, in this model as well, but agents, activities, places and events are divided into individual entities from the description of the material. Another difference to the traditional model is that the archive does not have to be a collection of documents produced and received by a single records creator. In the new model the archive can, and will, contain documents created and received by several agents. This means organisational mergers, spin-offs and name

changes will not hamper the management of the archive. In the traditional method the organisation and management of an archive have always started from the records creators and required an early decision on which documents belong to which records creator. This is a very inflexible system, and making changes to it afterwards has been very difficult and cumbersome. In the new model everything is based on links which are easy to unlink and relink.

The effects of the new data model on the management of materials

Elka warmly welcomes the new data model for two reasons. Firstly, the new archival description makes it significantly easier for end customers to search for and discover information. Thanks to the new description model, the content of the archive materials take a key role for the first time. Until now, the purpose of archival description has been to describe the structure of the archive, and therefore support the preservation of its value as evidence. In the new model, however, the focus is on the material itself, and on the improved efficiency of findability and use. Subjects, ontologies and classifications are used in a way that makes it easy and efficient for the customer to find the content they are looking for. In the traditional system the customer had to first know which organization, ie records creator, was responsible for the content. After this, the customer had to search for the correct document in the records creator's archive to possibly find the data they are looking for. In the OSA system that uses the Capture data model the customer can start searching for data directly, as it is no longer necessary to know which organisation has processed the information or in which archive the data is stored.

The second immediate benefit is related to the handling of archived documents in the archive regardless of the format, electronic or paper. With the new data model, the materials can be inventoried, screened, grouped and described on the fly. In the traditional model a lot of time had to be spent for getting to know the history of the records creator and for limiting the archives, since each archive could only have one records creator, and it was not possible to mix documents from different records creators. The new model allows the technical repairing and cataloguing of documents on the fly, even when detailed background data is not available. The agents – previously known as records creators – places, events and activities are separate entities that can be linked to the materials whenever a corresponding relationship is discovered. This is a very useful feature, since knowledge of the archive increases as the materials are sorted. In the old model all information had to be available at the very outset of the sorting.

OSA to replace six old systems

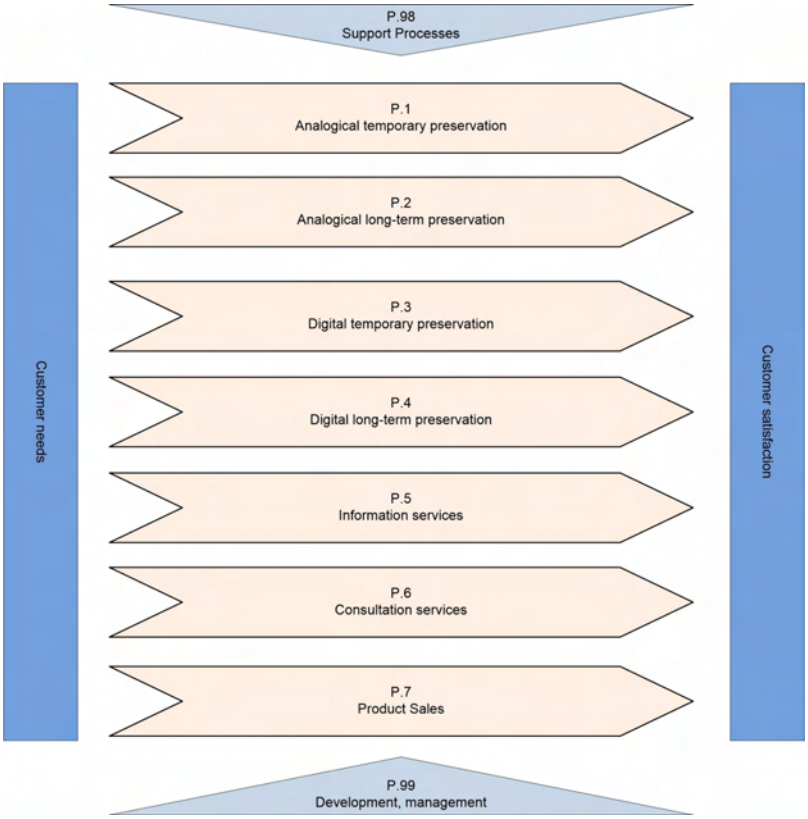
Elka currently uses six different systems to manage its archives. The Elma archive database, launched in 2002, contains descriptions and catalogue information on records creators, archives, document series and archiving units. Therefore, it is a very traditional archive database. Elma also contains four different registers for the so-called special materials, such as photos, drawings, maps and audiovisual records. In addition, Elka has a separate ElkaD application for presenting digitised sound records and moving pictures.

It is of course not sensible to divide materials and metadata like this across several different systems. In the worst case, the same materials must be catalogued several times, such as when photographs are described into the archive database as archive units, and also as individual photographs into the photograph register. This also makes searches by end customers cumbersome, since they must use the search features of several systems. The maintenance of several systems is also expensive for us as an organisation. Fortunately, all six data repositories can be managed by the OSA system at one go.

The uses of the new information system

During the first phase of the Capture project in 2011 Elka created service descriptions that brainstormed, listed and described the services that we want to offer our customers in the future (Alm & Strömberg 2013, 6–7). The implementation of these services served as a yardstick, when the specifications of the new IT systems were created in the next phase of the project. At that time, we did not know that we would be able to use the specifications documents as soon as we were. The reason was that this Open Source Archive project started in the summer of 2012 while the Capture project was still ongoing, and one of the most important projects of OSA has been to create a service archive system that conforms to the Capture specifications.

Picture 3 describes the main processes that were specified and described in the Capture project. These include permanent and temporary storage of both analogue and digital materials. These services are aimed at our corporate customers. In addition, the new IT system helps us in information services, consultation services and product sales. The image of an information service bases on its user interface which must be clear, easy-to-use, helpful and accessible. It must also have features for sharing information, such as links to social media. We also encourage our users to log in to the service. When users have logged in, they can be given access to materials that are subject to copyright, and they can also order materials to be sent to the research room.



PICTURE 3. Elka's main processes

Many archives have constructed separate digital archives for managing electronic materials, while they also keep using traditional archive databases for managing paper archives. We want to reject this idea, since the key issue for both analogue and digital materials is metadata. Some metadata are associated with digital objects, while others are not. Moreover, a document can have various forms, such as an original born-digital document, a printed and signed version and even a digital document created by scanning the printed and signed version. It would be utterly pointless to manage each of these forms in a separate information system.

We are currently living in a hybrid stage while we move away from analogue archives towards digital archives. In practice, documents are still created both in paper and in digital formats for decades. A modern archive therefore contains analogue, born-digital and digitised documents. The new open source archiving system will enable us to manage all these versions efficiently.

Summary

Elka participates in the Open Source Archive project to be able to have perhaps the most modern archive information in the world, or at least one that suits us best. From the perspective of an archive manager, the most exciting part is not the software or hardware itself, but the data model inside the system. The contextual description data model helps our end customers in many ways, and us archive managers, too. Contextual archival description will revolutionise the way data is searched for in an archive. For the first time in history the data content itself will override the document type as a search criterion.

As an archive the OSA system enables us to provide faster, more efficient and cheaper services to our corporate customers. Compared to our current IT system, OSA will enable us to launch a full- fledged electronic archiving service, including the management of materials at the customer's site or rented premises. Moreover, the compliance with international standards enables us to share our materials in The National Digital Library and other portals managed by a third party.

References

Alm, Olli & Strömberg, Janne 2013. *Capture-projekti – loppuraportti*. Mikkeli: Elka.

Alm, Olli & Lampi, Mikko 2014. *Flexible data model for linked objects in digital archives*. Archiving conference in Berlin 2014.

Arkistojen kuvailu- ja luettelointisäännöt 1997. Helsinki: Arkistolaitos.

Lybeck, Jari et al 2006. *Arkistot yhteiskunnan toimiva muisti. Asiakirjahallinnon ja arkistotoimen oppikirja*. Helsinki: Arkistolaitos.

Working together with Fedora Commons: sustainable digital repository solutions

Chris Awre

A persistent URL for the final article: http://www.darchive.fi/osa/

Abstract

Fedora Commons (most often shortened to just 'Fedora') is open source digital repository software that is maintained by an active community of institutional contributors from around the world. The ongoing development of Fedora is coordinated through the non-profit DuraSpace Foundation. Institutions join DuraSpace as members, contributing a fee that provides core staffing and strategic planning for Fedora as well as DSpace and associated services. DuraSpace works closely with the community-led Fedora Leadership Group, which comprises members of senior Fedora users. It should be noted that there is no link to the Fedora Linux distribution, which is entirely separate.

How did Fedora get to this point? It started as a computer science project at Cornell University in 1996 based around the question, 'If you could design a system to manage any type of digital content, what would it look like?'. Fedora development continued on this project basis until 2003, when the University of Virginia expressed interest in developing the system into a stable production version of a digital repository platform. Fedora has steadily attracted interest for a broad range of digital content management use cases around the world. Why is this? Fedora was selected by the University of Hull for the following reasons:

- It was designed to scale up
  - The amount of digital content is only going to grow
- It was designed to be content agnostic
  - We don't know what content types will need managing in the future
- It was designed to be based on open standards
  - Facilitating interoperability between systems

- It was designed to support the management of related items and describe the connection between them
    - Very little content lives in isolation
- It was designed to support the durability and preservation of digital content
    - To help digital content be usable into the future

Whilst other systems might provide some of this capability, only Fedora can provide it all. These principles have been maintained through the recent development of Fedora 4, which is a complete re-write of the codebase but one that is very much aiming to extend Fedora's functionality, not change it. Fedora 4.0 beta is currently available and aimed at new Fedora users: Fedora 4.1 will be released later in 2015 for existing Fedora users to migrate to.

Fedora has been applied for a wide range of purposes: for collections of texts, collections of images, datasets, audio and video collections, as well collections made up of combinations of these. Key to making use of Fedora for these different purposes has been to model what you want to do with the content. Fedora makes you think about this, and forces serious consideration of how the content will be managed, both now and in the future. This can be hard, but the effort is worthwhile and increases the likelihood of sustainability for the collection.

Fedora has often been described as a system that can support digital preservation. It certainly has functionality built in that enables content to be durable, for example, creating checksums, auditing actions against objects, and enabling the repository to be re-built if it becomes corrupted. Fedora does not provide everything for preservation, though, and relies on interaction with other systems to carry out preservation actions. Fedora is best thought of as a system that enables preservation to take place. In this context, a number of sites have implemented Fedora as the basis of a preservation solution, interacting with other tools and services as required.

Fedora has always been very flexible as part of enabling the characteristics described above. Because of this the system has never had a default end-user interface, as having this would immediately place boundaries around how the system operated. Fedora adopters have created their own interfaces to suit the content they are managing. This is an operational burden, but one that those making use of Fedora have found useful so they can apply their own solution. It has, though also prevented some organisations from adopting Fedora, as they are concerned about the level of technical effort required to maintain the solution.

In recent years two major open source initiatives have sought to make adopting Fedora more straightforward: Hydra and Islandora. Both have created frameworks that make generating interfaces for creating, reading, updating and deleting content much easier, using tools based on Ruby on Rails and Drupal, respectively. Both are seeing considerable take-up globally, and both are seeking to build communities of their own to sustain the developments. Commercial partners providing services based on the software can help support adoption. Solutions for all the same use cases as Fedora have been developed, but in a way that makes sharing those solutions more straightforward than before.

What next for Fedora? As has been mentioned, Fedora 4 development is ongoing, and will reach maturity during 2015. There is a lot of interest and emphasis on the use of RDF as the basis for storing digital content objects within the system, although XML-based content can also be managed as well. This adoption of linked data as the basis for storing digital collections promises to be valuable for the sustainability of the collections and allowing those collections to be used in new ways. Fedora will continue to be a valuable asset itself for working with digital content for some considerable time.

The harmonization of digital contents: benefits and requirements

Mikko Lampi & Aki Lassila

Introduction

Information governance in distributed environments is challenging due to the complexity and diversity of contents, data sources and standards (W3C 2009). However, harmonization enables coherent access and discovery for the materials in information systems. Therefore, the availability of harmonized data promotes the discovery of useful information and relations within the data that might otherwise remain undetected. Furthermore, it improves interoperability and usefulness of the information. Harmonization can be achieved by utilizing transformations, mappings and by linking metadata via ontologies, vocabularies or other linked data.

As generally known, information can be indexed and further analyzed. Language and entity identification supports natural language processing and the understanding of language specific properties. Indexing provides fast access and additional ways, such as faceting and geospatial analysis, to discover, access and visualize the information. Harmonized metadata can be linked and exposed as public or private linked data. The use of native linked data technologies enables the efficient exploitation of information and open data (Bizer et al. 2009).

Searching has become more than just using a web search engine such as Google or Bing. Searching is now associated with discovery platforms with full-text and natural language processing capabilities, also including features such as visualizations, facets and mashups. In addition, usability and user experience are very important factors in search and access, and the platforms should support complete machine-readability and data interoperability. Moreover, the trustworthiness of the data sources are important aspect.

This article demonstrates the concepts and technologies of harmonization with two projects: Open Source Archive (OSA) and Capture. OSA is a project carried out by Mikkeli University of Applied Sciences (Mamk) and funded by

the European Region Development Fund. The project started in June 2012 and ends in December 2014. The primary objective of the OSA project is to find and develop open source tools and solutions for digital archives. Its key features include archival materials and lifecycle management, long-term preservation, ingest, search and access. The OSA software is based on well-known open source software. Later in this article, OSA is used to refer to the digital archive software unless stated otherwise. The OSA project is based on Capture which was a data modeling and digital archiving system definition project of Central Archives for Finnish Business Records (Elka) and Mamk. It was carried out during 2011 - 2012. The primary deliverables from Capture were the concept of a harmonized metadata model and the specifications for a modern and flexible digital archive system.

Extracting data from digital contents

The first step in harmonizing digital contents is to extract the metadata and the file content into a machine-readable form. Extraction requires that each file format has a compatible parser. There are easily tens of formats for rich text documents, audio, moving image, pictures and other valuable digital contents. Each format requires a parser library which can extract file's technical metadata, embedded descriptive metadata and some, or all, of the actual data content. For archival purposes one must know the significant information for the specific format in order to correctly preserve it. Different tools provide different technical outputs which need to be mapped and processed before forwarding the information for harmonization and indexing. After the initial extraction the data is in usable form, but by no means, harmonized or even normalized.

A widely used extraction solution is Apache Tika. It can be used to extract information from documents and to detect the language automatically. Tika can identify the file and automatically select a suitable parser, if known. Automation can be achieved by integrating Tika or other data extraction solutions with indexing engines. Tika is widely used in archival software developed in Mamk.

Indexing information

Indexing is necessary for efficient access to huge amount of textual data such as extracted contents of rich text documents. Usually the index itself is a binary format data store. It does not replace or make obsolete the original data but supports its usage. Indexing is required to enable feasible and efficient

processing of time consuming tasks such as full-text search and certain analysis processes.

The basic principle in indexing is the same across different technical solutions. Databases and other data stores can be indexed for faster read operations and information retrieval. Write operations become slightly slower, but the performance gain is multiple. This is because write operations are usually done less often while read operations are more or less continuous. Search engines use indexes to rapidly find relevant information based on the search terms and to then return the objects from the data store. While most of the operations could be completed without indexes, they would often be very inefficient. The performance difference is even more drastic if the data is read from a file system or from disks instead of memory.

Furthermore, full-text indexing is very useful for accessing unstructured digital content. It enables full-text search which users are used to when using search engines such as Google. Other benefits of full-text indexing include statistical information based on the indexed terms and their respective hit rates. Full-text search is discussed in more detail later in this article.

One of the most used indexing solutions is Apache Solr which is also used in OSA. In addition to indexing Solr provides search features and tools for simple statistical analysis. It can be extended with various plugins such as information extraction with the previously mentioned Apache Tika.

Language processing is a critical part of full-text indexing. It provides an accurate and valid identification of terms and entities. Some languages, such as Finnish, have inflected forms and therefore require the basic forms of words to be determined. This can be very problematic without vocabularies. There are also other entities, such as proper nouns, which need to be identified and indexed correctly. In some cases specific entities need to be removed to protect privacy or confidentiality. In the OSA project, Apache Solr was used in combination with the Voikko library for accurate Finnish language indexing and queries. Due to the nature of the Finnish language, Voikko also includes an extensive vocabulary in addition to the grammatical rules. Voikko is an open source project and used in projects such as LibreOffice. The integration of Voikko and Solr was developed as open source by The National Library of Finland as part of its Finna project (http://www.kdk.fi/en/public-interface/software-development).

In addition, indexed terms can be linked with ontologies or vocabularies to gain formal definition and better interoperability. For example, indexed Finnish place names could be linked with the national spatio-temporal ontology SAPO. This way, the information would be more usable than utilizing the unnormalized terms.

Metadata harmonization

Metadata harmonization is a process consisting of multiple steps, both technical and non-technical. The motivation for harmonization includes interoperability and feasibility (e.g. Nilsson 2010). While a lot of entities described are related to humanistic sciences, for instance, rather than technical, the information systems require structured and machine-readable data. The results provide new or better services for consumers and better understanding of the materials.

Different fields and industries have specified their own metadata standards to support their contents and activities. For example, MARC21 (http://www.loc.gov/marc/) is widely used in libraries and LIDO (http://lido-schema.org) in museums. Most of the standards have in common that they support well the specific metadata and objects, but are not intended for managing information systems or information exchange. LIDO, for example, covers all kinds of museum objects such as art, architecture, cultural history, natural and the history of technology. LIDO enables the creation of normalized records for museum context. These records can be further enhanced by providing ontology linking. Semantic records can then be shared with other systems and environments.

Metadata interoperability is one of the primary reasons for metadata harmonization. Interoperability requires that metadata records are machine-readable and compatible with each other. Dublin Core Metadata Initiative defines metadata interoperability as the ability of two or more agents, such as information systems and software components, to exchange metadata so that the interpretation remains consistent with the original context and information (Nilsson 2011).

Interoperability means that normalized records conform to metadata models which can then be mapped to ontologies, vocabularies and other metadata models which can be internal models or metadata standards. OSA uses mapping as the primary method for harmonizing metadata. The basic principle is to map various input formats into an internal umbrella metadata model. The OSA ingest workflow first harmonizes the data into a master data model which is then used to generate the index. Benchmarking was carried out to compare functionality and the concepts of existing projects such as Finna (See NDL 2014). OSA and Finna serve different purposes, but the reasons for harmonization are the same. They both need to ingest diverse contents and provide access and management in a coherent manner. It is not feasible to implement different user interfaces, application logic and user experience for each kind of data.

During the Capture project additional reasons for interoperability were identified. Firstly, it was confirmed that in archiving there is a need for digital archive and repository services, preferably hosted as SaaS. This kind of solution was built as part of the OSA project. The approach had a lot of different files, metadata, standards and formats put into one system which all the tenants share. It has a single core repository, Fedora Commons, which manages the content and the metadata. Fedora can manage all the files and metadata formats as separate streams but it can turn out as a complexity creep and hard to manage environment. It is more efficient to harmonize as much as possible. (Lampi & Alm 2014.)

An umbrella metadata model, known as the Capture model, was designed to tackle the harmonization challenges. The Capture model was designed to be compatible with several national and international metadata models such as Dublin Core, SFS 5914, JHS 143 and SÄHKE2 (Alm & Strömberg 2013). It can be extended to support other standards and custom metadata definitions if needed. Because of the extent of the unified model, its smaller piece can be defined as a content model for various content types. Each content type is fully compatible with the main model. In addition, metadata values can be links to ontologies and vocabularies. Content described with the Capture model form a linked data network which can be private, public or a hybrid. (Lampi & Alm 2014.)

Furthermore, an important lesson learned is that a harmonized model cannot dictate too many restrictions. The umbrella model must support all kinds of needs and provide a coherent internal harmonization framework. Restrictions like cardinality and locale-based settings need to be applied in interfaces pulling and pushing the data. In OSA mappings and transformations are an integral part of the architecture. Because OSA is a multi-tenant environment each organization has its own set of mappings which binds the data to user interfaces and APIs. Each mapping is also archived so that the original meaning and knowledge for reading it are preserved. The mappings can be executed technically with any suitable transformation method such as XSLT. This way harmonization is a lossless and two-way process.

The harmonization process should be automatic, which means that the data models and interfaces have to be machine-readable. It is achieved by providing sufficient technical information for processing the data models, metadata and contents. The data itself has to be structured or otherwise machine-readable. Finally, there need to be APIs for data operations. The APIs can be public or private, and a public API can be used to deliver non-public content.

Finally, harmonization is not all about technology. A very important factor is communication between all the parties involved. Understanding the context and meaning of the materials is essential in preserving it unaltered during

the process. The feedback and improvement process should be iterative and continuous.

Search and access

Search in this context is more than a textbox-based search engine such as Google or Bing. It is a combination of a discovery portal, browsing catalog, a recommendation and curation engine and a technical platform. Search is a method of finding interesting records and objects in possibly huge data sets, but selected sources.

Traditional search engines find information in various or even unknown sources based on the search terms provided by the user. The results can be in any format depending on the source materials. The metadata is often very limited or truncated. The algorithms and indexes are good, but all else depends on luck. With digital archives, repositories and other kinds of collections, one cannot afford Google-like results If it is not on the first page, it probably won't be found. And, if Google cannot find it, it doesn't exist at all.

Search in OSA provides a highly configurable search interface. It includes a familiar full-text search, and depending on the configuration, multiple search fields and pre-fetched facets such as temporal, spatial and content types, and some visualizations for these. Furthermore, it is possible to estimate the accuracy and number of the search results before actually rendering the results. Of course, full-text search can be used the same way as the Google or Bing search.

Search is performed against the selected and reliable source. As mentioned above, OSA is not a web search engine. Instead, it finds contents in its index which contains public contents as well as restricted and confidential contents. By default OSA searches materials based on the user information such as organization, roles and access rights. If no user information is found, it will only search public materials for a specific organization. OSA is not a portal and it searches content in one organization at a time. Each organization requires its own search page. Currently, there is no cross-organization search, but it is technically possible to build. Full-text search can understand languages, identify basic forms, synonyms and other entities. The search also covers the contents of rich text documents such as PDFs.

Search results are scored and returned with harmonized and standardized metadata and in easy-to-read and understandable format. Scoring is determined by algorithms and help to determine the quality of the search results. Rich metadata enables a configurable result page and additional methods for refining the results. OSA has a completely configurable results view which can be easily adapted to various result types. Each organization can define

the significant metadata which is displayed automatically. The search view can show different amount of information based on if the user is logged in or not, and depending on the roles and access rights. Harmonization makes it possible to use consistent search terms and facets to search and filter digital content. Therefore, OSA provides access to diverse metadata records and files in a coherent manner. It supports storing the original metadata as additional information.

Metadata records, file previews and such can be displayed for the search results. All the data available in the index can be used for searching and exposed as a facet. Facets are valuable before and after the search. Before the search, facets can provide suggestions and completion features and help to choose search terms that will return meaningful results. After the search, they can help to profile the results and filter the records. In addition, OSA provides download and management options according to roles.

The technical solution for search in OSA is Apache Solr. The front-end and search logic is built as a multi-tier web application. The front-end is based on earlier development done in digital archive projects and services. A lot of effort has been put on usability. The development model is based on the agile methodologies and emphasis is put on listening to the feedback from participants.

Content discovery

OSA demonstrates discovery and analysis by utilizing the object network created by Fedora Commons. Each entity archived or stored in OSA is a compound object consisting of multiple data streams. Fedora Commons uses a specific stream to store each object's relations in RDF/XML format. Relations are then indexed to a resource index which is a RDF database. By default Fedora Commons 3 uses Mulgara which can be queried e.g. with the SPARQL language. Objects in the RDF database form a linked data network. OSA supports relations of any kind between the objects, but currently only Dublin Core relations and a content model definition are used. The relations network enables analysis on how entities are related to each other and how distant the relations are. Another example is the archival hierarchy catalog which can be built automatically and dynamically from the isPartOf relations.

Discovery was found useful in the Capture project when planning how the existing object network could enrich new objects during the ingest and the description processes. The basic concept is that an object gains partial or complete contextual information from the surrounding linked objects such as

agents, places, events and actions. These contextual objects can be formalized via ontologies or vocabularies. (Lampi & Alm 2014.) This improves the description speed, and information duplication is minimized. Enrichment can take place during ingest or access depending on the need. The principle is the same regardless of the timing. The process can be automatic or controlled by users. It can add information to the object's metadata or just modify the index leaving the original object unaltered.

Summary

Based on the experiences and lessons learned in the case projects and benchmarking, it can be said that harmonization is an integral part of search, discovery and access. Depending on the source materials harmonization can require extraction and normalization before indexing can be done. The current trend in repositories and archives is towards digitalization which causes fast growth in the amount and in the diversity of digital content. The experience and research done in memory organizations can help commercial companies to identify, classify and harmonize their materials. This is because challenges are more or less similar with every kind of content regardless of the owner organization.

In addition, new tools related to big data, analysis and data mining could add value to the existing data that is stored in the information systems of memory organizations. However, in order to utilize new technologies and methods the data must be in good condition. Regarding usability, there are different aspects in content usability: machine-readability, context awareness and user experience to name a few. Furthermore, content analysis could be used in completely new applications such as data-based leadership and decision making applications. Statistical information about index usage could prove useful in developing services which consume the harmonized content.

This article covered a lot of development and research done at Mamk, but also benchmarked related projects such as Finna. Many of the concepts and topics are merged and result in new features and added value to the existing applications. Development started in OSA is not completed, when the project comes to an end. The work requires constant development and evaluation of the latest research and tools in the field. This is the kind of dialog with the community that has been going on during the past couple of years, and it is also the right direction for future collaboration. Still, there is plenty of room for future research and development. These include managing the information overload, automatic curation and the preservation of important information and experiences, as the generations before us have done.

References

Alm, Olli & Strömberg, Janne. 2013. Summary of Final Report for Capture Project. WWW document. http://www.elka.fi/useruploads/files/Summary.pdf. Retrieved 20 October 2014.

Bizer, Chrtistian, Heath, Tom & Berners-Lee, Tim. 2009. Linked Data - The Story So Far. International Journal on Semantic Web and Information Systems.

NDL. 2014. NDL Public Interface Project at The National Library of Finland. WWW document. https://github.com/KDK-Alli. Retrieved 24 October 2014.

Lampi, Mikko & Palonen, Osmo. 2013. Open Source for Policy, Costs and Sustainability. Archiving Conference Proceedings (Vol. 2013).

Lampi, Mikko & Alm, Olli. 2014. Flexible Data Model for Linked Objects in Digital Archives. Archiving Conference Proceedings (Vol. 2014).

Nilsson, Mikael. 2010. From Interoperability to Harmonization in Metadata Standardization: Designing an Evolvable Framework for Metadata Harmonization.

W3C. 2009. WWW document. Improving Access to Government through Better Use of the Web. W3C Interest Group Note 12 May 2009. http://www.w3.org/TR/egov-improving. Retrieved 24 October 2014.

Web-application development in the Open Source Archive project

Liisa Uosukainen

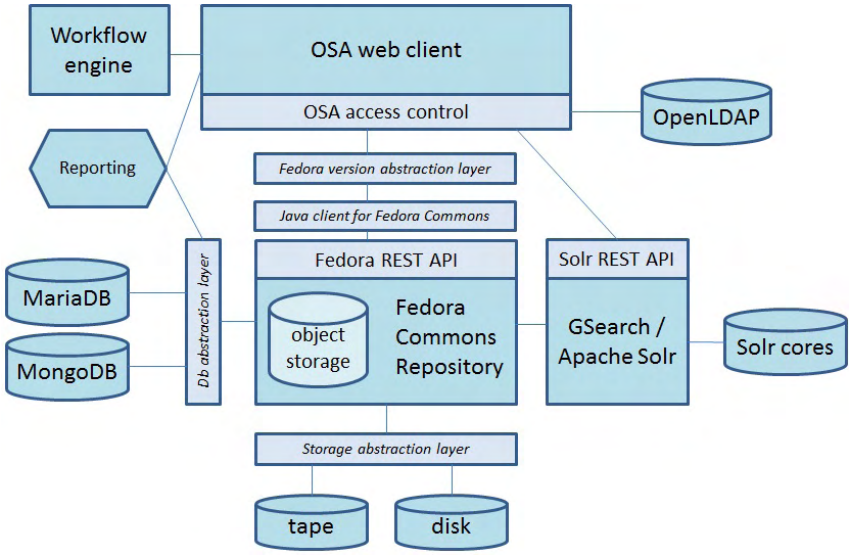
Introduction

Open Source Archive (OSA) is an EU funded project that is carried out by the Mikkeli University of Applied Sciences. One major part of the OSA project was to develop a service-oriented archive solution for preserving, managing, and providing access to digital content. Open Source Archive (OSA) project started in May 2012, and runs until the end of December 2014. As partners OSA had archives, software vendors, service providers and educational institutions.

This article introduces the technical aspects of the OSA application development focusing on the components involved in the OSA application. In addition to technology selection and technical design we put efforts into making user interfaces user-friendly. A couple of studies concerning service design, data visualization and user-centered methods in the user interface design were made as bachelor theses during this project. The results of these studies provided us with the guidelines for user interface development.

Technology solutions and OSA web-application implementation

The OSA service-oriented archive is a web-based application that provides user-friendly access to the implemented features such as ingesting, searching, preserving, and destroying of documents. The OSA application is written in Java. It is composed of software components which are used via common APIs (Lampi, M., Palonen, O., 2014). The architecture of the software components is described in Picture 1 below. While the focus was to develop a sustainable solution providing long-term access and retention, we selected open source software components in order to become more independent of software vendors.



PICTURE 1. OSA software architecture and system components

Object-oriented programming and Model View Controller (MVC) design were the key principles of the implementation. We intended to create as modular architecture as possible where each module would operate independently. Modular design enabled us to develop and test each component separately. Moreover, it reduced the internal complexity of the application system. The replacement of one module with a different implementation is possible without having to change the rest of the source code, since the implementation is hidden from the other components using it.

We designed the OSA application to support multi-tenancy. A single instance of the OSA application serves the needs of multiple tenants, i.e. organizations. Each organization can customize several features of the OSA application to meet their needs. Some examples for this are listed below.

- It is possible to define properties for UI concerning the organization specific title, CSS, a logo, and header files.
- Metadata creation is made flexible. The system supports extendable and configurable types of digital objects that are preserved including schemas. Organizations can configure their user interfaces and metadata fields for specific digital objects individually. Search and indexing properties are also configurable covering visible fields in a search form, columns in the result layout, and facet fields.
- The system allows organizations to create their own records management plan. Relating features are available for system administrators and records managers who have full access to digital content.

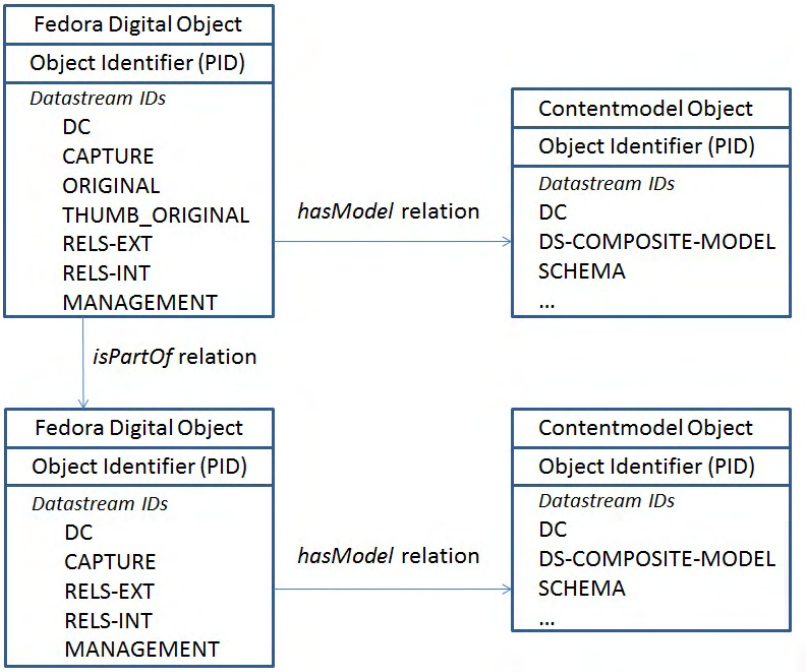
Customization for each tenant is done via organizations' XML configuration files; the modification of the source code is not required.

Fedora Commons Repository

The OSA archiving application is based on Fedora Commons Repository which is an open source repository platform. Fedora Commons Repository was originally developed at the Cornell University and the University of Virginia. Nowadays the Fedora community is supported by the DuraSpace nonprofit organization. The software enables storage, management and long-term access to digital resources. Fedora is written in Java, and it offers access via REST or SOAP interfaces (Fedora Documentation, 2014).

Fedora is designed to be flexible, supporting all types of digital content. It implements Content Model Architecture (CMA) which ensures that objects comply with the model (Razum, M. et. al., 2009). We defined content models for collections, documents, images, moving images, audio recordings, maps, and so on. Each content model has an XSD schema that is used for validating digital objects before ingesting them into the Fedora repository. The OSA application shows new content models automatically as types of digital content, when content models are stored into the Fedora repository. The OSA application creates dynamically ingest forms for these type of specific content models as well. Ingest forms containing descriptive metadata can be customized by organizations. In order to adjust the metadata into standard terms the OSA application maps customized elements into the Capture format.

The Fedora digital object in a repository is the combination of the following components: PID, object properties and datastreams. Original files as well as the metadata are *datastreams* in an object model. Digital objects are stored in a Fedora repository in the FOXML (Fedora Object XML) format which encapsulates object properties and datastreams. The picture below describes the structure of digital objects in the OSA archiving application. Each object has a persistent identifier (PID) that is used as a reference to the object and the datastreams. The OSA application stores metadata into DC (Dublin Core) and CAPTURE datastreams. The actual content is in ORIGINAL datastream. The RELS-EXT datastream links an object to the content model and describes relationships to the other digital objects. The RELS-INT datastream, in turn, can be used for describing internal relationships in the object context. The OSA application maintains a special MANAGEMENT datastream in order to manage access control and preservation according to the preservation plan of the OSA application.



PICTURE 2. Digital objects in OSA archiving system

Fedora provides an option to index RELS-EXT and RELS-INT datastreams automatically to the RDF-based Resource Index (Fedora Documentation, 2014). Fedora uses Mulgara as a default RDF database. The properties of RELS-EXT and RELS-INT datastreams can be queried through the Resource Index Search Service (RISearch) interface by using the RDF query languages, SPARQL or iTQL. The OSA application uses RISearch in order to find the hierarchical structures of objects. The most common search is to find the isPartOf relation of objects when generating collection views for browsing purposes.

Fedora keeps track of changes in the object repository. It maintains an audit trail that contains information on every change made to the object. Content versioning is done at datastream level, not at object level. If a datastream property is set to be versionable, every modification of a datastream leads to its new version (Razum, M., Schwichtenberg, F., Fridman, R., 2007).

The OSA application uses MediaShelf to communicate with Fedora Commons Repository 3.6.x. MediaShelf provides Java client for Fedora REST API, but it only supports Fedora 3.x versions. We have the intention to upgrade the OSA application to support the Fedora 4.0 version as soon as a new Java client

supporting Fedora 4.0 is published. Fedora 4.0 has various improvements to 3.x versions, such as performance and usability improvements. At the end of this project Fedora 4.0 is heading towards a production release. The third beta release of Fedora 4.0 was published in August 2014.

Apache Solr

Apache Solr is an open source search platform that indexes the repository content. Fedora 3.x supports Solr indexing by using the GSearch (The Fedora Generic Search Service) component, which has been developed by the Technical University of Denmark. GSearch automatically updates Solr indexes via JMS (Java Messaging Service), when Fedora objects are created, modified or purged.

We used multicore Solr setup and installed two SolrCore instances: one for indexing metadata and another for indexing content. When each core has its own search indexes, we can define schemas for them and keep the size in control. The OSA application utilizes Solr's faceting component which divides search results into categories and shows their final numbers. The metadata fields for faceted search are configurable in the OSA application. Solr with the Voikko plugin - available in GitHub by the National Library's KDK project - brings more usability enabling natural language search, spell checking and grammar checking, for Finnish.

Databases

The OSA application as well as Fedora requires a relational database to store data related to certain application functionalities. We selected MariaDB, a community driven, open source database management system, for that purpose. MariaDB is a successor of MySQL, nowadays owned by Oracle. The OSA application uses MariaDB to store simple information in a database containing a few tables.

For more complex and large datasets, we installed MongoDB. It is a popular NoSQL document database that supports the storage of documents in the JSON, CSV, or TSV format (The MongoDB 2.6 Manual, 2014). MongoDB offers both a free community version and a paid-for enterprise version. MongoDB turned out to be a convenient place to store diverse forms of metadata of digital objects. It stores different kinds of metadata groups in the same collection providing an index on any attribute.

OpenLDAP

OpenLDAP software is an open source implementation of Lightweight Directory Access Protocol (OpenLDAP, 2013). It provides directory services for user authentication and role based access control (RBAC). According to Sandhu, Ferraiolo, and Kuhn (2000) the basic concept of RBAC is that permissions are assigned to roles, users are assigned to roles and users acquire permissions by being members of roles.

The OSA application's role determines a set of permissions that users, assigned to that role, are allowed to perform. The permission defines an access right to the collection and an access right level to the content. The access right level defines the functions that can be performed, such as capability to read or modify or delete metadata and/or digital records. It is possible to define a specific publicity level, such as public, confidential, restricted, for the permission which implies that all the digital records with lower publicity level are available. Users who have permission to manage roles in organizations can create roles for various purposes. Many roles can be assigned to one user and one role may have several permissions. RBAC ensures that data is accessible and reusable to both the information professionals and the consumers at an appropriate permission level.

Workflow engine

The archiving processes of digital documents are likely to be automatized whenever it is appropriate. Automated processes were carried out by utilizing a workflow engine that has been developed as a bachelor thesis in the OSA project. The workflow engine supports microservice architecture where processes are modeled as actions, called microservices, and combined into workflows (Kurhinen, H., Lampi, M., 2014). Workflows are configured in an XML-based configuration file. New services are easy to create and to combine to workflows without having to modify the workflow engine itself.

We created a set of micro services for pre-ingest workflow and ingest workflow. We also provided a user interface where users can ingest several files at the same time and the whole content will be handled in the same way during the ingest process. The pre-ingest workflow covers tasks such as virus checking, technical metadata capturing, checksum checking, and preview generation. The ingest workflow finally uploads digital objects to the archive. We designed a feature to trigger workflows automatically, when files are uploaded onto a certain directory. Still, there is a possibility to start workflows manually.

OSA development methods

The OSA project applied agile software development methods. According to Abbas, Gravell and Wills (2008) a development method is agile, when it is adaptive (welcomes the change and responds to feedback), iterative and incremental (software is developed in several iterations), and people-oriented (prefers communication within the team and with the customer). Agile development helped us to build the working software containing the most desired functionalities. The system requirements were already available. The previous Capture project produced the functional requirement specification for a modern archiving system (Alm, O. & Strömberg, J., 2013). Small features were selected for design, implement and test. In this way, we were able to respond to the feedback quickly by building features one small piece at the time.

Project partners' role

Project partners were involved to test the OSA application at a very early stage. The proto of the OSA application was available on a web site. We created organizations and usernames for the project partners for testing purposes and published new versions of the OSA application regularly giving the project partners a chance to evaluate the new features. The possibility to send feedback was made as easy as possible. Customers were able to provide feedback for the application and testing results by clicking the *send feedback* link in the OSA application. All reported bugs and improvement ideas were updated into a bug tracking system, called Mantis Bug Tracker, where they were prioritized against new features. In order to improve communication we had several feedback meetings, mostly with ELKA's (Central Archives for Finnish business records) records managers and other specialists who had been testing the OSA application.

All in all, gathering customer feedback throughout the project enabled our efficient software development. Feedback helped us to prioritize some features and to improve the usability of the OSA application. It was also a benefit that the application was tested with real-world customer data instead of unilateral test data generated by software developers.

Conclusions

Choosing the agile project management methodologies for this project turned out to be a suitable way of implementing the OSA application. Feedback from customers was included flexibly in the development process. In this way, we were able to concentrate on the features that were the most important to the

customers. When developing a persistent digital archive, there is a need for sustainable software development as well. We relied on selecting open source components that are developed by active communities, so that the continuing development and support should be guaranteed. The lack of documentation or outdated documentation especially concerning some installation and configuration instructions was a disadvantage. However many of the communities provide mailing lists that are good for information exchange. Our ambition is that there will be a community of OSA users that continue the work done in this project.

References

Abbas, Noura, Gravell, Andrew M. & Wills, Gary B. 2008. Historical Roots of Agile Methods: Where Did "Agile Thinking" Come From? In Proceedings of Agile Processes in Software Engineering and Extreme Programming: 9th International Conference, XP 2008, (Limeric, Ireland, June 10-14, 2008).

Alm, Olli & Strömberg, Janne 2013. Capture-projekti - loppuraportti, Mikkeli: Suomen Elinkeinoelämän Keskusarkisto.

Anderson, Richard 2012. Digital Object Storage and Versioning in the Standford Digital Repository, Digital Library Systems and Services, Standford University Libraries and Academic Information Resources, Standford Digital Repository. PDF document. https://lib.stanford.edu/files/Digital Object Storage and Versioning.pdf. Retrieved 15 October 2014.

Fedora Documentation. 2014. DuraSpace. WWW document. https://wiki.duraspace.org/display/FEDORA/All+Documentation. Retrieved 17 October 2014.

Kuć Rafał. 2013. Apache Solr 4 Cookbook (2nd ed.), Packt Publishing Ltd., Birmingham.

Kurhinen, Heikki & Lampi, Mikko 2014. Micro-services based distributable workflow for digital archives. In Proceedings of Archiving 2014, (Berlin, Germany, May 13-16, 2014).

Lampi, Mikko & Palonen, Osmo 2014. Open Source for Policy, Costs, and Sustainability. In Proceedings of Archiving 2013, (Washington DC, USA, May 13-16, 2013).

MediaShelf Fedora-client. 2013. MediaShelf. WWW document. http://mediashelf.github.io/fedora-client/site/0.7/index.html. Retrieved 14 October 2014.

OpenLDAP. 2013. OpenLDAP Foundation. WWW document. http://www.openldap.org/. Retrieved 2 October 2014.

Razum, Matthias, Schwichtenberg, Frank & Fridman, Rozita 2007. Versioning of Digital Objects in a Fedora-based Repository, In German E-Science Conference 2007, (Baden-Baden, Germany, May 2-4, 2007). PDF document. http://www.escidoc.org/media/docs/ges-versioning-article.pdf. Retrieved 21 October 2014.

Razum, Matthias, Schwichtenberg, Frank, Wagner, Steffen & Hoppe, Michael 2009. eSciDoc Infrastructure: A Fedora-Based e-Research Framework. Research and Advanced Technology for Digital Libraries. 13th European Conference, EDCL 2009 (Corfu, Greece, September 27 – October 2, 2009). Volume 5714, pp. 227-238.

Sandhu, Ravi, Ferraiolo, David & Kuhn, Richard 2000. The NIST Model for Role Based Access Control: Toward a Unified Standard. In Proceedings of 5th ACM Workshop on Role Based Access Control, (Berlin, Germany, July 26-27, 2000).

The MongoDB 2.6 Manual. 2014. MongoDB Inc. WWW document. http://docs.mongodb.org/manual/. Retrieved 19 September 2014.

Archival UI – an old relic or modern and novel?

Anssi Jääskeläinen

Participatory design methods and novel techniques are utilized as a part of designing information technology solutions for everyday life. However, when focusing on the archival field things are, frankly speaking, vice versa. Information technology is used to transfer the analog world into the digital world as it is. This article focuses on the digital archive user experience from the UX professional point of view. Still, the presence of the paper archives as well as the influence of the paper archive users is strong. It is not uncommon to encounter a new digital archive that is used by browsing through a hierarchical tree. The problem, however, is not the actual system or its design, it is the users. While the paper-generation users don't see anything wrong in browsing a hierarchical tree or filling in paper-like metadata forms, the IT generation users are likely to start laughing at such a system or bursting into tears. There is a huge contradiction between the requirements of these user groups. This article is about the ongoing work around bringing these groups and their requirements closer to each other by utilizing relic as well as novel methods in finding the happy medium. The ultimate intention has been, and still is, to provide the best possible user experience (UX) for the users of digital archives in spite of their old habits. The work behind this article has been partially conducted by a bachelor thesis writer Tytti Vuorikari who currently works for a company called AIE design.

Introduction

Countless number of electronic devices, applications or online services have flopped due to various reasons. These might include, but are not limited to, defective operation, missing support, usability issues, a fuzzy user interface (UI), illogical workflow or general bad user experience (Cooper 2004). One of the most well-known examples is Windows Vista which was loaded with new functionality and candy UI features, such as Aero. Still, it only received the status of being hated among its users. Undoubtedly, Vista had a lot of technical and usability issues, but still the biggest problem from the authors' point of view was its difference to the de-facto standard of that time, Windows XP.

Resistance to change is something that needs to be taken into account when something new is pushed into a traditional field, such as archives. While from the theoretical UX point of view the change might be excellent, the traditional end users might still hate it due to their different habits and viewpoints.

The key of the research done in the OSA project with the aid from other experts has been in the rapid and agile development methods and in thorough consideration of user requirements and wishes. This work has been going on successfully around digital archives for more than ten years now.

The starting point situation was, however, simpler than in the traditional archives, since the burden of the paper archives was already gone. However, generally speaking the transition from paper to digital is far from being straightforward. Some of the clients of our archive service have taken the journey from the analog archive to the digital one, and this is where we participated in the process. The transition is carried out in cooperation with paper-generation users, IT generation users, designers and developers with the help of a UX professional.

The rest of this article describes some of the methodologies that have been used to enhance the user experience of our digital archive solution. For example, a new core version of digital archive UI has been designed and developed by utilizing these modern and novel methods.

UX in the traditional field

Participatory design methodologies, User Centered Design (UCD), personas, interaction design, the honeycomb model, user experience lifecycle, etc. are just some of the available procedures for enhancing the end user experience (Lowdermilk 2013). These are used every day for example in modern IT, software, electronic and vehicle development. Digital archiving business is a very dedicated area, although it is based on information technology. Roughly speaking, when dedication towards a certain operational area increases, more IT vendor oriented the solution will be. This is due to lack of competition which leads to a vendor monopoly situation and the powerlessness of the customers. One of the most illustrative examples is the library section in Finland where only one software vendor has existed until very recent times.

When simplified enough, any digital archive is based on a technical backend and an end user frontend. Technical backend such as a storage center, tape library and service management are controlled by the IT administrator. These backend control interfaces and their functionalities are not meant to be used by the end users, who are bound to use the client interface provided.

Developing an archival client interface with archival controls including, pre-ingest, ingest, preservation, access, and delivery are IT dependent tasks into which we have been placing our development efforts. The intention has been to develop our client interface as user-friendly as possible, even though there hasn't been any considerable competition in the field. Now this work conducted has been starting to pay off, since the competition is starting to get stronger mainly due to public clouds. These clouds, including even Facebook, are often considered by average Joes as safe places for storing information. This is happening especially among the younger generation. While Facebook or Google Drive can be regularly backed up places for storing information, they are still far away from being true digital archives.

Basic platform or personalized environment

The technological backend and the core of the client interface are the same for all clients and users. However, the design has always been modular so that it can easily be extended or personalized according to clients' requirements. The normal procedure with clients has always been the following: Before the service agreement the client is met and their requirements, needs and wishes are recorded in an agreement. Before the start-up situation it is checked together with the client that the agreement terms and requirements are met. A direct feedback channel between the client and development team is also active during the customership.

During the past few years, we have tried to take one step further to offer even more tangible digital archive user experience to the clients in spite of their background. This has been done by utilizing different service design methods such as service concepts, customer profiling and service blueprints in designing the next archival client interface release. Through these methods we intended to receive answers to questions such as: How do users process information? What kind of work do users do, and how and why do they do it? What information is relevant and what is not? Do all users have the same needs? Attention was especially paid to rationalizing the different user needs and user workflows. Users were also empowered by letting them participate in the research and development process. For example, they had an opportunity to conceptualize their own tools with a co-creation method. The next subsections will introduce the most utilized methods.

Survey

A survey that formed the basis for this work was conducted in 2012 and the results were published in a conference (Jääskeläinen 2012). The survey was about collecting user data from archival users to be able to model the true ar-

chival user experience. The survey had a total of 42 respondents from the different areas of digital and analog archives. Ease of use, information retrieval, information reliability, system stability and the speed of the search were found to be the most important features in a digital archive, while for example, the aesthetics of the UI was not found to be an issue.

This is something that modern IT generation users won't understand: How is it even possible to do work with a UI that is illogical, mainly text-based, doesn't support Copy and Paste and contains no Undo functionality, just to mention few examples? When the results from of survey were compared against another survey that was conducted among university students and UX professionals, the differences were notable (Jääskeläinen 2012 and 2013). This is one of the reasons why these novel methods were brought to the field of digital archiving as well.

Shadowing with interviews

Shadowing is commonly used when something that already exists needs some development. This ethnographic method offers detailed information about the interaction process between the user and the task, whether analog or digital. By utilizing the shadowing method with the users of analog archives their operational models and habitual ways can be uncovered, which can then be modeled as a digital workflow. In general, shadowing provides information about different usage and behavioral patterns between the user and the target (Stickdorn and Schneider 2012). Together with the interviews, shadowing allowed us to create user profiles and design drivers that have been very helpful in creating the design and functionality for the new archival control UI. Furthermore, shadowing gave us good starting points for designing the transition process from the analog to digital archives.

The shadowing method revealed the behavioral patterns of users, but to find out why users behaved how they did during the shadowing, focused interviews were utilized. The basis for the interviews was a core set of questions which were enhanced with additional questions in case the interviewed had something to add. A total of six people were interviewed during this phase of the development work. The basic problem with the interviews is the Hawthorne effect described by Jones (1992), which basically means that the behaviors of the interviewed persons change to meet the expectations of the interviewer. The possibility of this effect was minimized by making the interview situation as "uninterviewlike" as possible. The interview process was continuous throughout the shadowing day and did not start until the subject of the shadowing seemed to be comfortable with the ongoing situation. This way, the true needs, desires, points of view and the motivations of the users were discovered.

Personas and other methods utilized

Information gathered from the shadowing and interviews were utilized in producing user profiles which were then visualized with the personas method. Personas are realistic snapshots of the possible end users and helpful for developers to recognize why something has to be done in a certain way, and in communication with the end user via persona (Cooper 2004). From the information gathered, five different personas with user profiles were created and utilized. One of these personas is presented as a persona card in Picture 1. The persona card is in the Finnish language, but its purpose is to only give an idea of how personas are built.

PICTURE 1. An example of a persona card by Tytti Vuorikari

User profiles were utilized further in generating a written service story on how different users are using the service. The stories created were then used as a basis for the service blueprint which is a visualization model of the service in a chronological order. Finally, a simple UX expert evaluation has been carried out throughout the development, and suggestions were delivered to the developers.

Results and conclusions

All of the described methods have been utilized, and the results gained are currently used in the development of the next major release of our archival control system, which is solely based on open source. The beta version of the archival control UI has been released and it can be accessed via the address osa.mamk.fi. The end user functionality is based on Google-like search and on faceted focusing on the search results.

However, for the sake of paper generation users, a possibility to browse the tree structure has been maintained in the archive. Many features which were identified as important by archive users, such as a basket and advanced tree structure browsing, were also included in the release. With the feedback from the current beta testers, who are also clients of the present system, the new control UI and its functionality will be further developed. By the end of this year, a modern and novel production release should be launched.

References

Cooper, Alan 2004. The inmates are running the asylum. Sams Publishing

Jones, Stephen R. G. 1992. Was There a Hawthorne Effect?, American Journal of Sociology, 98, 3.

Jääskeläinen, Anssi 2011. Integrating User Experience into Early Phases of Software Development. Doctoral Dissertation, Lappeenranta University of Technology.

Jääskeläinen, Anssi 2012. Rationalizing the concept of user experience in digital preservation, Conference proceedings Archiving 2012, Copenhagen

Lowdermilk, Travis 2013. User-Centered Design, CA: O'Reilly

Sticdorn, Marc & Schneider, Jakob 2012. This is service design thinking. BIS Publishers

Open Source in Mamk's IT education

Paula Siitonen & Matti Juutilainen

Background

Education has been digitized like any other field of industry. Nowadays we talk about online learning or elearning or even MOOCs. In online learning study materials are available in digital form, students receive guidance via the internet, course participants can discuss the topics in virtual meeting rooms and so on. Students are allowed to learn in some other place than in a school building. They stay at home doing exercises, carry out their practical training at work or gather together with schoolmates to study collaboratively. Students can use the internet for searching information and web meeting tools for discussing matters with teachers or fellow students remotely. They can submit their exercises to teachers via internet tools. All this requires different kinds of ICT tools: connections, applications and devices.

In Finland the Ministry of Education finances the education. In addition, not even universities are allowed to collect any fees from students studying to complete a full degree. Even students coming from abroad can study without fees in Finland. Because the financial situation is becoming more challenging in Finland, too, the Ministry of Education is reducing the funding for universities. In consequence, cheaper ways have to be found to provide education of good quality more efficiently and by saving in expenses. One possibility where costs can be reduced is information technology. Although it seems that there is need for even more technical solutions, savings can be achieved through finding and choosing the cheaper ones, for example, alternatives that base on open source.

This article aims at discussing such cheaper, open source, alternatives from the viewpoint of IT education and especially Mamk, ie Mikkeli University of Applied Sciences. In addition, it introduces how the soon ending project called OSA (Open Source Archive) where Mamk played a major role, was related to teaching within Mamk's degree programme of Information technology. The article first briefly introduces current topics involving open source and education. These include MOOC, open source, open application programming interface and open data. After that follows a short description of Mamk

as an institution and open source user in the field of IT education. The last part concentrates on the experiences of lecturer Matti Juutilainen who mainly participated in the OSA project as a teacher, and describes below how he has used the open source approach in teaching, and how his students were involved in OSA.

Current topics of education and open source

Starting with MOOC (Massive Open Online Course) it is quite a new model of learning online. MOOC is an online learning course where hundreds of students self-organize their participation. These courses are often free of charge and there are no prerequisites. Several universities around the world have delivered their own MOOCs especially due to the possibilities to lower the expenses, perhaps even at the cost of risking the quality of education. Even a couple of Finnish universities offer some MOOCs. Among others, Yan and Powell (2013) have written about the phenomena of MOOC and the trends towards greater openness in higher education as well as its implications for educational institutions.

When turning to discuss open source software, it is usually understood as software of free charge, including source code. This simple description gives a good overview, but there is a detailed definition made by Open Source Initiative which outlines the precise criteria the open source software must meet. The criteria address distribution terms and modification, for instance. (Open Source Initiation, 2014).

Open source is more of a style of developing and modifying applications. Anyone can use, copy, modify and distribute open source applications without license fees. According to its definition it is possible to choose applications freely, so that the decision is not dependent on certain operating systems or technologies. Open source is based on the idea that also new versions of open source software are available. Open source software is developed in most cases in developing communities where bugs are found and repairs made faster. That makes open source applications more reliable and of higher quality.

Moving on to open application programming interfaces, there are numerous applications already made for some specific purpose. It would save some money, if it would be possible to reuse some applications as parts of other information systems. To make this happen, software has to be able to communicate with other software. Also hardware has to be easy to connect with other hardware. Open application programming interface (OAPI) means that applications can interact with each other, they can change data or send requests to each other. Especially public administration has begun to demand

that information systems should provide open interfaces (API Suomi, 2014 & National Land Survey of Finland, 2014). A further advantage of OAPIs is that applications based on them do not need to be open source applications and the data transferred between this kinds of applications does not need to be open data.

To finish this first part of the article with open data, the Open Knowledge Foundation Network (OKFN) has been founded to promote openness (Open Knowledge Foundation, 2014). Open data is data that is available and accessible. It should be downloadable over the internet and in a convenient and usable form. Open data is reusable and has to be modifiable and mixable with other data sets.

Public administration in Finland has actively promoted operations contributing to openness. The Open Data Programme, led by the Ministry of Finance, focuses on eliminating obstacles that prevent the use and reuse of public data and on helping public administration to make data available (Ministry of Finance, 2014). In addition, for example the city of Helsinki has opened data for the free use of citizens (Helsinki Region Infoshare, 2014), and so has the National Land Survey in Finland (NLS). There is a self-service for loading maps on the NLS website. There are also services available for accessing some restricted set of data from the NLS servers by specific spatial data applications (National Land Survey of Finland, 2014). From the viewpoint of public administration, the target in opening data for use is to make the new innovative open data services to serve the citizens better.

IT Education and open source at Mamk

When continuing to introduce Mamk, Mikkeli University of Applied Sciences, it provides education in seven fields of study. There are two campuses, one in Mikkeli and one in Savonlinna. Although there are still lessons in classrooms, more and more learning takes place remotely, too, at home or at work, for instance. Learning that is independent of time or place has been on the increase for a couple of years.

Mamk provides two degree programmes that concentrate on ICT contents: The degree programme of Information technology focuses on network technologies, hardware and software engineering with some media content. The programme is concerned with data security and maintaining and upgrading server systems. Also some programming is included in the compulsory studies. The objective of this degree programme is to educate innovators for the future internet. The degree programme of Business Information Technology, in turn, concentrates more or less on programming. The objective is to learn how to make customer-tailored applications with modern and appropriate

programming environments and technologies. The main idea of both degree programmes is that students concentrate on certain skills thoroughly rather than on learning the overall basics of several principles, theories and technologies.

Mamk's online studies are supported by the browser-based Moodle online learning environment. Moodle was chosen because it provided an open interface with a low price, was widely used among universities and because of the maintenance services available. Although it is an open source application, there are support services available.

Although Mamk uses quite a lot of open source software in its IT education, this does not mean that Mamk would be committed to open source only. Instead, open source is the alternative chosen when it suits the purpose best. For example, databases are studied completely with the open source database MySQL and the first programming language is the open source programming language PHP. Another open source programming language, Java, is a natural continuation and used to learn programming techniques more thoroughly. In mobile programming Android techniques are used, and there are Linux-based systems available among operating systems. In addition to the open source applications and tools already listed, Mamk has quite low-priced university licenses for Microsoft applications. There is a set of operating systems available, office applications, programming languages and database systems, also for students. That is why these commercial tools are quite widely used among students and teachers as well, especially among non-experts in ICT.

To end this section and to move on to the article's last part, ie Matti Juutilainen's experiences, it is worth highlighting that Mamk is quite an active player in the field of RDI as well. Especially long-term preservation, filing, data management and recording systems have been under research. The latest activities have related to open source tools in the long-term preservation and data management developed in the OSA project. This project has been significant in many respects. One of the most remarkable advantages is that the researchers and teachers participating in the project have had the opportunity to become more thoroughly familiar with open source applications. And, we should not forget the students either. They benefited from the project as well, as Matti concludes below.

Experiences of open source and OSA by lecturer Matti Juutilainen

As already introduced above, I participated in the OSA project mainly in the role of a teacher. I also participated in some practical tasks in the project, such

as in configuring the Blade cluster with CentOS and OpenNebula with the Sunstone user interface. But most of the time I concentrated on linking the project with the courses and students at Mamk where I teach courses related to PC computer technology, operating systems, networking technologies and enterprise server environments. Participating in the OSA project provided me with new ideas that I have been able to use especially in my server related courses. Additionally, I collected numerous useful topics for exercises, labs, homework assignments, student projects and bachelor's theses.

Using the open source approach in teaching has given many benefits for me. As no financial investment is needed, we can freely evaluate and use different open source tools. For example, there are several widely used Linux operating system distributions to select from, and they can be used to demonstrate students what really happens behind the scenes on the operating system and application levels. The inner workings of the operating system and the tools are not hidden, as is the case with commercial options.

Although Linux is quite rare in desktop use, I have found it useful to teach students the basic ideas of its usage and shell already starting from their first year of studies. That gives a good background for their later studies, as for example, the command line interface of Cisco's networking device closely resembles the Linux shell interface. During the third year studies we reach the enterprise server environment where Linux is extensively used.

The OSA project has provided me with a valuable opportunity to update and extend my knowledge of some open source related topics, especially archiving solutions and long-term storage of information. As a teacher I have been able to link the project work and my teaching together. Through the project I have got numerous practical examples that I can share with my students. The project has been a good source of new ideas and experiences to be utilized in my courses. Digital archiving and long-term storage are very timely topics in the Mikkeli region and teaching these topics to our students may help some of them to actually to get employed by some local companies.

There are many specific challenges in the long-term preservation of digital information, especially with large amounts of data. These challenges include being independent of commercial players that may not meet the demands in the long run. Technically speaking, the challenges result for example from the redundancy and capacity requirements of the environment, such as the clustering of servers and using an external storage for large-scale databases, as well as from the software requirements, such as open file formats that can be converted to some other format when needed. I have found that many open source archiving tools and platforms are currently under rapid development,

which makes them a challenge to utilize as updates come frequently and new features are added constantly. Also documentation commonly lags a couple of steps behind, which makes the installation and configuration process tricky.

During the environment and development tasks of the OSA project we also identified many possible topics for our students of the international degree program of Information technology. Most topics were suitable for bachelor's theses (15 ECTS credits) or for smaller project studies (5 ECTS credits), as we usually could control the amount of work by including or excluding topics to work with. These topics related, for example, to studying NoSQL database solutions and measuring their performance in a clustered environment and selecting a suitable NoSQL implementation based on the requirements of the OSA project. Further topic examples included researching the options to migrate virtual machines from other virtualization platforms to the Linux KVM, studying high-performance clustering techniques in Linux, unrestricted streaming techniques that allow the real-time conversion of data from the archive format to a more suitable format for the end user and so on.

For newcomers, such as my students, these project topics proved to be very demanding. The environment itself caused the first challenge, as the students needed to connect to the servers remotely over a VPN connection and the starting point was a fresh installation of CentOS Linux in a virtual machine. The first part was to get familiar with the environment and possibly to deploy more virtual machines, and the second part was to start to configure them as a cluster. Additionally, the students did not have very much previous knowledge on the specific project topic, for example the NoSQL databases, and they needed to get familiar with the topic itself before they could actually come up with any background requirements for the server/operating system configurations.

Although I have been teaching students server hardware and software related technologies as well as Windows and Linux server administration, they seemed to be quite overwhelmed, when they needed to start working in a real environment with quite an abstract definition of the goals, rather than exact, step-by-step instructions for what to do and where. In practice, the projects required the students to start Googling around to gain knowledge of the topic, then to be able to select – maybe after evaluation – the best option to use in the implementation and then to follow the documents found in the internet to actually make the configurations work.

We clearly found out during the projects that the documentation for many archiving-related open source tools lags behind the development. This is most probably because the tools are not widely used, and the number of develop-

ers is rather small. Additionally, the development pace is fast and the features are constantly updated. Because of this, it was challenging for the students to follow the documentation as the latest software versions required different configurations compared with the documentation.

Still, despite the fact that the project topics were demanding for the students, the projects gave them a good insight on the current demands of working life in that particular field. Participating in a timely topic and making demanding configurations with inadequate instructions required the students to challenge their current knowledge and to learn the topic more deeply as successfully completing the configurations required understanding of the underlying platform.

The project work can be recognised as a valuable experience in the students' resumes. It can later help them to find a job in the IT field. As academic work is usually public, the student reports may also help others to gain ideas and experiences on the topic, or to find help when facing a problem. One potential outcome of the student projects is also real contribution to the current open source development by reporting – and even fixing – errors in the current implementation, adding patches to improve or increase functionality and by helping to improve the documentation.

References

API Suomi 2014. WWW document. http://apissuomi.fi. Referred 20.10.2014.

Helsinki Region Infoshare 2014. WWW document. http://www.hri.fi/en/. Referred 20.10.2014. Referred 21.10.2014

Ministry of Finance 2014. Open Government. WWW document. http://www.vm.fi/vm/en/05_projects/0238_ogp/index.jsp. Referred 20.10.2014.

National Land Survey of Finland 2014. WWW document. http://www.maanmittauslaitos.fi/en. Referred 20.10.2014.

Open Data 2014. WWW document. https://www.opendata.fi/en. Referred 20.10.2014.

Open Knowledge Foundation 2014. WWW document. https://okfn.org/. Referred 20.10.2014.

Open Source Initiative 2014. Open Source Definition. WWW document. http://opensource.org/osd. Referred 20.10.2014

Paikkatietohakemisto 2014. WWW document. http://www.paikkatietohakemisto.fi/geonetwork/srv/en/main.home. Referred 20.10.1014.

Yan Li, Powell Stephen 2013. MOOCs and Open Education: Implications for Higher Education. Centre for Educational Technology & Interoperability Standards. University of Bolton. PDF document. Updated http://publications.cetis.ac.uk/wp-content/uploads/2013/03/MOOCs-and-Open-Education.pdf. Referred 20.10.2014.

Open movement – a megatrend of our time

Noora Talsi & Mikko Lampi

Opening up the society

Openness is one of the megatrends of our time. Governments, educational organizations and businesses face new challenges when societies are moving towards openness and transparency. Opening data and joining the movement of open knowledge require especially changes in attitudes and ways of thinking. In a democratic society citizens and non-governmental organizations are entitled to know what their government is doing. To do that, they must be able to freely access governmental data and information and to share this information with other citizens. (Open Data Handbook 2012.)

Digitalization is a tool for openness and for sharing the information about open data and how to access it. The scale of openness is larger than ever before. The public discussion and hype around the open movement sometimes forget that most of the data that has now become open and available has actually always been open. Only the access has been the much more limited than now. Being able to get information about public announcements and decision-making has so far required citizens' active interest and effort. Now the data is available regardless of time and place. Open data is not enough, if no one can find it. Therefore, the definition of the term open data nowadays also conveys the meaning that data is not only open, but accessible as well.

Making data accessible and usable requires good software development and digital services. Making lists of opened data in a PDF format and publishing them on the government's website is of course open and accessible data, but it is not usable. Finding, understanding and being able to use information is important for citizens. Citizens' and the nation's collective memory is stored in archives. Still, data storage is not an archive – even though it may seem like one for average customers. Cloud services providing free storage space for customers have made storing information easier than ever before. As Anssi Jääskeläinen (2014) wrote earlier in this publication: "These clouds, including even Facebook, are often considered by average Joes as safe places for storing information. This is happening especially among the younger

generation. While Facebook or Google Drive can be regularly backed up places for storing information, they are still far away from being true digital archives".

It is quite obvious that opening and digitalizing great amounts of data face us with new challenges of information management. Saving information is easy, but it is much more difficult to understand the value of the data and to place it into a right context. Information management should cover the whole life cycle of information. When producing new data the ultimate question should be: How is the information managed, preserved and finally destroyed? Digital archiving gives us possibilities for reliable preservation, management, migration and for destroying information. The world is depending more on correct and high-quality information than ever before, which also emphasizes the role digital archiving.

The Finnish government has committed to open data and the six biggest cities in Finland have created a network devoted to open the data that they produce. This is remarkable for the civil society, democracy and the citizen contribution to decision-making. By opening up data, citizens are enabled to be much more directly informed and involved in decision-making. One of the key values of open data is transparency. Transparency is not only about accessing the original information, it is also about sharing and reusing it. For citizens transparency also means the right for reading and contributing to the decisions that local or national authorities are making. (Open Data Handbook 2012.)

Citizens have more opportunities to influence their social surroundings and the society than ever before. Citizens also have more tools to manage the information about themselves, and above all, more ways to know what information have been gathered about themselves and by whom. My Data is a movement towards human-oriented management of personal data. In the center of the My Data movement are the individuals who have the right to access and to obtain the personal data concerning themselves. (Poikola et al. 2014.)

Open data as an opportunity for business and education

A transparent and democratic society for citizens is one side of the open data movement, the other side is the opportunities for business. Opening data provides a significant way of creating new business: software, applications as well as services. Open data as a business opportunity becomes obvious when comparing the weather data and the related business in different continents.

In the US weather is three times bigger business than in the EU. The US weather data have had open access policy for decades. Most of the publicly funded weather data is available inexpensively and as widely as possible, whereas European weather data has had national monopolies to control the data to recover the costs of its collection and creation. As a result, most of the weather data is chargeable, and related business by third parties is practically impossible. (Turtiainen 2014.)

Archives and other memory organizations store remarkable and valuable contents in either analog or digital form. In future, technological advancements and better knowledge of opening data enable the large scale utilization of archival materials. However, there are a lot of questions, such as privacy and copyright, to take into consideration. Not all materials can be published and opened as such. Memory organizations are in a way forerunners in long-term information management and in making information available. The availability of archived materials could be very useful and valuable for business, decision-making, travel, culture and history. Finding and accessing information is essential in using digital contents, as Mikko Lampi and Aki Lassila wrote in their article on the benefits and requirements of harmonization of digital contents.

The demands of open data and knowledge and its digital distribution have reached educational organizations as well, as Matti Juutinen and Paula Siitonen introduced earlier in their article about education at Mamk. MOOCs – Massive Open Online Courses – have quickly spread globally and the discussion around MOOCs has also started in Finland. At the same, the role of teachers or lecturers has been re-evaluated. In online teaching the lecturer is no longer a teacher in the traditional sense, but more of a facilitator of learning – someone who makes the learning experience possible. The other big issue around MOOCs is the openness itself. The MOOC movement, originated in the Ivy League universities, follows the principles of open access and freedom. However, the context of expensive private universities in the US is just totally different from the free public universities and universities of applied sciences in Finland. The democracy and equal rights to higher education that MOOCs have made possible in the US have always been the key value in the Finnish higher education. In Finland higher education is seen as a civil right that everyone should be able to access regardless of wealth. And what comes to the US, it has been estimated that 80 per cent of the students who have completed MOOCs are already enrolled in universities. In other words, the MOOCs have tempted only few new students. (Hiidenmaa 2013.)

Besides teaching, the other important question related to openness in universities and universities of applied sciences is publishing. The results of scientific research that is publicly funded have been traditionally published in esteemed scientific journals with limited access. The latest scientific information is accessible only for the university elite. The movement of open access publishing has been created as a counter force by individualists and open-minded researchers. Our principle with this publication was to publish it online and as an open access volume.

Open source is about freedom of information: most commonly understood as access to software's source code, but it is also about the freedom of speech applied to software (See Open Source Initiative). In the Open Source Archive (OSA) project it was recognized that building a sustainable digital archive software is possible only with open software. Open movement on a global scale has confirmed this decision. Memory organizations such as the Library of Congress of the United States have adopted open source as a replacement for proprietary solutions. Furthermore, the public sector is moving towards open information systems. A prime example is the National Data Exchange Layer which is based on the Estonian X-Road, released as open source. Open source based information systems and software are independent of vendor lock-ins, specific technologies and proprietary components. These are essential criteria when building e.g. digital archives which need to last for decades or for centuries. However, open source in itself is not a silver bullet. It enables a shared and transparent ecosystem, but requires knowledge and continuous effort.

The article of Osmo Palonen described above the three periods of digital archiving at Mikkeli University of Applied Sciences this far. We are proud to add the next, the fourth chapter. Our aim is to establish a research center for digital information management at Mamk in 2015. The center will focus on the applied research on digital preservation and information management. Openness is a key value of our new research center, too.

So why should we care so much about openness? Simply because the alternative is quite scary. The world without access to data and knowledge does not support transparency, democracy or civil rights. A society where all the information is in the hands of few does not work well and is not a good environment for people, business nor organizations. That is why, openness is important also for those who do not have a special interest in the open knowledge movement, not to mention open source coding. Valuable information such as cultural history and a nation's memory need to be stored in trustworthy systems, be it a digital archive or something else. Trust is created by openness, transparency and independence.

References

Hiidenmaa, Pirjo 2013. Jos vastaus on mooc, mikä on kysymys? If MOOC is the Answer, What is the Question?]. Helsinki: University of Helsinki.

Open Data Handbook Documentation 2012. Open Knowledge Foundation.

Open Source Initiative. Open Source Definition. WWW document. http://opensource.org/osd. Retrieved 1.11.2014.

Poikola, Antti, Kuikkaniemi Kai & Kuittinen, Ossi 2014. My Data – Johdatus ihmiskeskeiseen henkilötiedon hyödyntämiseen My Data – An Introdction to Human-centered personal data]. Helsinki: Liikenne ja viestintäministeriö.

Turtiainen, Heikki 2014. New environmental services utilizing open data. Avoin Suomi.

MIKKELIN AMMATTIKORKEAKOULU
MIKKELI UNIVERSITY OF APPLIED SCIENCES. MIKKELI. FINLAND

PL 181, SF-50101 Mikkeli, Finland. Puh.vaihde (tel.vx.) 0153 5561

Julkaisujen myynti: Tähtijulkaisut verkkokirjakauppa, www.tahtijulkaisut.net. Julkaisutoiminta: Kirjasto- ja oppimisteknologiapalvelut, Kampuskirjasto, Patteristonkatu 2, 50100 Mikkeli, puh. 040 868 6450 tai email: julkaisut@xamk.fi

MIKKELIN AMMATTIKORKEAKOULUN JULKAISUSARJA

A: Tutkimuksia ja raportteja ISSN 1795-9438
Mikkeli University of Applied Sciences, Publication series

A: Tutkimuksia ja raportteja – Research reports

A:1 Kyllikki Klemm: Maalla on somaa. Sosiaalinen hyvinvointi maaseudulla. 2005. 41 s.

A:2 Anneli Jaroma – Tuija Vänttinen – Inkeri Nousiainen (toim.) Ammattikorkeakoulujen hyvinvointiala alueellisen kehittämisen lähtökohtia Etelä-Savossa. 2005. 17 s. + liitt. 12 s.

A:3 Pirjo Käyhkö: Oppimisen kokemuksia hoitotyön kädentaitojen harjoittelusta sairaanhoitaja- ja terveydenhoitajaopiskelijoiden kuvaamina. 2005. 103 s. + liitt. 6 s.

A:4 Jaana Lähteenmaa: "AVARTTI" as Experienced by Youth. A Qualitative Case Study. 2006. 34 s.

A:5 Heikki Malinen (toim.) Ammattikorkeakoulujen valtakunnalliset tutkimus- ja kehitystoiminnan päivät Mikkelissä 8. – 9.2.2006. 2006. 72 s.

A:6 Hanne Orava – Pirjo Kivijärvi – Riitta Lahtinen – Anne Matilainen – Anne Tillanen – Hannu Kuopanportti: Hajoavan katteen kehittäminen riviviljelykasveille. 2006. 52 s. + liitt. 2 s.

A:7 Sari Järn – Susanna Kokkinen – Osmo Palonen (toim.): ElkaD – Puheenvuoroja sähköiseen arkistointiin. 2006. 77 s.

A:8 Katja Komonen (toim.): Työpajatoimintaa kehittämässä - Työpajojen kehittäminen Etelä-Savossa -hankkeen kokemukset. 2006. 183 s. (nid.) 180 s. (pdf)

A:9 Reetaleena Rissanen – Mikko Selenius – Hannu Kuopanportti – Reijo Lappalainen: Puutislepinnoitusmenetelmän kehittäminen. 2006. 57 s. + liitt. 2 s.

A:10 Paula Kärmeniemi – Kristiina Lehtola – Pirjo Vuoskoski: Arvioinnin kehittäminen PBL-opetussuunnitelmassa – kaksi tapausesimerkkiä fysioterapeuttikoulutuksesta. 2006. 146 s.

A:11 Eero Jäppinen – Jussi Heinimö – Hanne Orava – Leena Mäkelä: Metsäpolttoaineen saatavuus, tuotanto ja laivakuljetusmahdollisuudet Saimaan alueella. 2006. 128 s. + liitt. 8 s.

A:12 Pasi Pakkala – Jukka Mäntylä: ”Kiva tulla aamulla…” - johtaminen ja työhyvinvointi metsänhoitoyhdistyksissä. 2006. 40 s. + liitt. 7 s.

A:13 Marja Lehtonen – Pia Ahoranta – Sirkka Erämaa – Elise Kosonen – Jaakko Pitkänen (toim.): Hyvinvointia ja kuntoa kulttuurista. HAKKU-projektin loppuraportti. 2006. 101 s. + liitt. 5 s.

A:14 Mervi Naakka – Pia Ahoranta: Palveluketjusta turvaverkoksi -projekti: Osaaminen ja joustavuus edellytyksenä toimivalle vanhus-palveluverkostolle. 2007. 34 s. + liitt. 6 s.

A:15 Paula Anttila – Tuomo Linnanto – Iiro Kiukas – Hannu Kuopanportti: Lujitemuovijätteen poltto, esikäsittely ja uusiotuotteiden valmistaminen. 2007. 87 s.

A:16 Mervi Louhivaara (toim.): Elintarvikeyrittäjän opas Venäjän markkinoille. 2007. 23 s. + liitt. 7 s.

A:17 Päivi Tikkanen: Fysioterapian kehittämishanke Mikkelin seudulla. 2007. 18 s. + liitt. 70 s.

A:18 Aila Puttonen: International activities in Mikkeli University of Applied Sciences. Developing by benchmarking. 2007. 95 s. + liitt. 42 s.

A:19 Iiro Kiukas – Hanne Soininen – Leena Mäkelä – Martti Pouru: Puun lämpökäsittelyssä muodostuvien hajukaasujen puhdistaminen biosuotimella. 2007. 80 s. + liitt. 3 s.

A:20 Johanna Heikkilä, Susanna Hytönen – Tero Janatuinen – Ulla Keto – Outi Kinttula – Jari Lahti –Heikki Malinen – Hanna Myllys – Marjo Eerikäinen: Itsearviointityökalun kehittäminen korkeakouluille. 2007. 48 s. + liitt. (94 s. CD-ROM)

A:21 Katja Komonen: Puhuttu paikka. Nuorten työpajatoiminnan rakentuminen työpajakerronnassa. 2007. 207 s. + liitt. 3 s. (nid.) 207 s. + liitt. 3 s. (pdf)

A:22 Teija Taskinen: Ammattikeittiöiden ruokatuotantoprosessit. 2007. 54 s.

A:23 Teija Taskinen: Ammattikeittiöt Suomessa 2015 – vaihtoehtoisia tulevaisuudennäkymiä. 2007. 77 s. + liitt. 5 s. (nid.) 77 s. + liitt. 5 s. (pdf.)

A:24 Hanne Soininen, Iiro Kiukas, Leena Mäkelä: Biokaasusta bioenergiaa eteläsavolaisille maaseutuyrityksille. 2007. 78 s. + liitt. 2 s. (nid.)

A:25 Marjaana Julkunen – Panu Väänänen (toim.): RAJALLA – aikuiskasvatus suuntaa verkkoon. 2007. 198 s.

A:26 Samuli Heikkonen – Katri Luostarinen – Kimmo Piispa: Kiln drying of Siberian Larch (Larix sibirica) timber. 2007. 78 p. + app. 4 p.

A:27 Rauni Väätämöinen – Arja Tiippana – Sonja Pyykkönen – Riitta Pylvänäinen – Voitto Helander: Hyvän elämän keskus. ”Ikä-keskus”, hyvinvointia, terveyttä ja toimintakykyä ikääntyville –hankkeen loppuraportti. 2007. 162 s

A:28 Hanne Soininen – Leena Mäkelä – Saana Oksa: Etelä-Savon maaseutuyritysten ympäristö- ja elintarviketurvallisuuden kehittäminen. 2007. 224 s. + liitt. 55 s.

A:29 Katja Komonen (toim.): UUDISTUVAT OPPIMISYMPÄRISTÖT – puheenvuoroja ja esimerkkejä. 2007. 231 s. (nid.) 221 s. (pdf)

A:30 Johanna Logrén: Venäjän elintarviketurvallisuus, elintarvikelainsäädäntö ja -valvonta. 2007. 163 s.

A:31 Hanne Soininen – Iiro Kiukas – Leena Mäkelä – Timo Nordman – Hannu Kuopanportti: Jätepolttoaineiden lentotuhkat. 2007. 102 s.

A:32 Hannele Luostarinen – Erja Ruotsalainen: Opiskelijoiden oppimisen ja osaamisen arviointikriteerit Mikkelin ammattikorkeakoulun opiskelijaarviointiin. 2007. 29 s. + liitt. 25 s.

A:33 Leena Mäkelä – Hanne Soininen – Saana Oksa: Ympäristöriskien hallinta. 2008. 142 s.

A:34 Rauni Väätämöinen – Merja Tolvanen – Pekka Valkola: Laatua arvioiden. Mikkelin ammattikorkeakoulun ja Savonia-ammattikorkeakoulun tutkimus- ja kehitystyön benchmarking. 2008. 46 s. + liitt. 22 s. (nid.) 46 s. +liitt. 22 s. (pdf)

A:35 Jari Kortelainen – Yrjö Tolonen: Vuosiluston kierresyisyys sahatavaran pinnoilla. 2008. 23 s. (pdf)

A:36 Anneli Jaroma (toim.): Virtaa verkostosta. Tutkimus- ja kehitystyö osana ammattikorkeakoulujen tehtävää, AMKtutka, kehittämisverkosto yhteisellä asialla. 2008. 180 s. (nid.) 189 s. (pdf)

A:37 Johanna Logrén: Food safety legislation and control in the Russian federation. Practical experiences. 2008. 52 p. (pdf)

A:38 Teija Taskinen: Sähköisten järjestelmien hyödyntäminen ammattikeittiöiden omavalvonnassa. 2008. 28 s. + liitt. 2 s. (nid.) 38 s. +liitt. 2 s. (pdf)

A:39 Kimmo Kainulainen – Pia Puntanen – Heli Metsäpelto: Etelä-Savon luovien alojen tutkimus- ja kehittämissuunnitelma. 2008. 68 s. + liitt. 17 s. (nid.) 76 s. +liitt. 17 s. (pdf)

A:40 Nicolai van der Woert – Salla Seppänen –Paul van Keeken (eds.): Neuroblend - Competence based blended learning framework for life-long vocational learning of neuroscience nurses. 2008. 166 p. + app. 5 p. (nid.)

A:41 Nina Rinkinen – Virpi Leskinen – Päivikki Liukkonen: Selvitys matkailuyritysten kehittämistarpeista 2007–2013 Savonlinnan ja Mikkelin seuduilla sekä Heinävedellä. 2008. 41 s. (pdf)

A:42 Virpi Leskinen – Nina Rinkinen: Katsaus matkailutoimialaan Etelä-Savossa. 2008. 28 s. (pdf)

A:43 Kati Kontinen: Maaperän vahvistusratkaisut huonosti kantavien maiden puunkorjuussa. 2009. 34 s. + liitt. 2 s.

A:44 Ulla Keto – Marjo Nykänen – Rauni Väätämöinen: Laadun vuoksi. Mikkelin ammattikorkeakoulu laadunvarmistuksen kehittäjänä. 2009. 76 s. + liitt. 11 s.

A:45 Laura Hokkanen (toim.): Vaikuttavaa! Nuoret kansalaisvaikuttamisen kentillä. 2009. 159 s. (nid.) 152 s. (pdf)

A:46 Eliisa Kotro (ed.): Future challenges in professional kitchens II. 2009. 65 s. (pdf)

A:47 Anneli Jaroma (toim.): Virtaa verkostosta II. AMKtutka, kehitysimpulsseja ammattikorkeakoulujen T&K&I –toimintaan. 2009. 207 s. (nid.) 204 s. (pdf)

A:48 Tuula Okkonen (toim.): Oppimisvaikeuksien ja erilaisten opiskelijoiden tukeminen MAMKissa 2008–2009. 2009. 30 s. + liitt. 26 s. (nid.) 30 s. + liitt. 26 s. (pdf)

A:49 Soile Laitinen (toim.): Uudistuva aikuiskoulutus. Eurooppalaisia kokemuksia ja suomalaisia mahdollisuuksia. 2010. 154 s. (nid.) 145 s. (pdf)

A:50 Kati Kontinen: Kumimatot maaperän vahvistusratkaisuna puunkorjuussa. 2010. 37 s. + liitt. 2 s. (nid.)

A:51 Laura Hokkanen – Veli Liikanen: Vaikutusvaltaa! Kohti kansalaisvaikuttamisen uusia areenoja. 2010. 159 s. + liitt. 17 s. (nid.) 159 s. + liitt. 17 s. (pdf)

A:52 Salla Seppänen – Niina Kaukonen – Sirpa Luukkainen: Potilashotelli Etelä-Savoon. Selvityshankkeen 1.4.–31.8.2009 loppuraportti. 2010. 16 s. + liitt. 65 s. (pdf)

A:53 Minna-Mari Mentula: Huomisen opetusravintola. Ravintola Tallin kehittäminen. 2010. 103 s. (nid.) 103 s. (pdf)

A:54 Kirsi Pohjola. Nuorisotyö koulussa. Nuorisotyö osana monialaista oppilashuoltoa. 2010. 40 s (pdf).

A:55 Sinikka Pöllänen – Leena Uosukainen. Oppimisverkosto voimaannuttajana ja hyvinvoinnin edistäjänä. Savonlinnan osaverkoston toiminnan esittely Tykes -hankkeessa vuosina 2006–2009. 2010. 60 s. + liitt. 2 s. (nid.) 61 s. liitt. 2 s. (pdf)

A: 56 Anna Kapanen (toim.). Uusia avauksia tekemällä oppimiseen. Työpajojen ja ammattiopistojen välisen yhteistyön kehittyminen Etelä- ja Pohjois-Savossa. 2010. 144 s. (nid.) 136 s. (pdf)

A:57 Hanne Soininen – Leena Mäkelä – Veikko Äikäs – Anni Laitinen. Ympäristöasiat osana hevostallien kannattavuutta. 2010. 108 s. + liitt. 11 s. (nid.) 105 s. + liitt. 11 s. (pdf)

A:58 Anu Haapala – Kalevi Niemi (toim.) Tulevaisuustietoinen kehittäminen. Hyvinvoinnin ja kulttuurin ammattikorkeakoulutuksen suuntaviivoja etsimässä. 2010. 155 s + liitt. 26 s. (nid.) 143 s. + liitt. 26 s. (pdf)

A:59 Hanne Soininen – Leena Mäkelä – Anni Kyyhkynen – Elina Muukkonen. Biopolttoaineita käyttävien energiantuotantolaitosten tuhkien hyötykäyttö- ja logistiikkavirrat Itä-Suomessa. 2010. 111 s. (nid.) 111 s. (pdf)

A:60 Soile Eronen. Yhdessä paremmin. Aivohalvauskuntoutuksen tehostaminen moniammatillisuudella. 2011. 111 s + liitt. 10 s. (nid.)

A:61 Pirjo Hartikainen (toim.). Hyviä käytänteitä sosiaali- ja terveysalan hyvinvointipalveluissa. Tuloksia HYVOPA-hankkeesta. 2011. 64 s. (pdf)

A:62 Sirpa Luukkainen – Simo Ojala – Antti Kaipainen. Mobiilihoiva turvallisen kotihoidon tukena -hanke 1.5.2008–30.6.2010. EAKR toimintalinja 4, kokeiluosio. Loppuraportti. 2011. 78 s. + liitt. 19 s. (pdf)

A:63 Sari Toijonen-Kunnari (toim.). Toiminnallinen kehittäjäkumppanuus. MAMKin liiketalouden koulutus Etelä-Savon innovaatioympäristössä. 2011. 164 s. (nid.) 150 s. (pdf)

A:64 Tuula Siljanen – Ulla Keto. Mikkeli muutoksessa. Muutosohjelman arviointi. 2011. 42 s. (pdf)

A:65 Päivi Lifflander – Pirjo Hartikainen. Savonlinnan seudun palveluseteliselvitys. 2011. 59 s. + liitt. 6 s. (pdf)

A:66 Mari Pennanen – Eva-Maria Hakola. Selvitys matkailun luontoaktiviteettien, Kulttuurin ja luovien alojen Yhteistyön kehittämismahdollisuuksista ja -tarpeista Etelä-Savossa. Hankeraportti. 2011. 29 s. + liitt. 18 s. (pdf)

A:67 Osmo Palonen (toim.). Muistilla on kolme ulottuvuutta. Kulttuuriperinnön digitaalinen tuottaminen ja tallentaminen. 2011. 136 s. (nid.) 128 s. (pdf)

A:68 Tuija Vänttinen – Marjo Nykänen (toim.). Osaamisen palapeli. Mikkelin ammattikorkeakoulun opetussuunnitelmien kehittäminen. 2011. 137 s.+ liitt. 8 s. (nid.) 131 s. + liitt. 8 s. (pdf)

A:69 Petri Pajunen – Pasi Pakkala. Prosessiorganisaatio metsänhoitoyhdistyksen organisaatiomallina. 2012. 48 s. + liitt. 6 s. (nid.)

A:70 Tero Karttunen – Kari Dufva – Antti Ylhäinen – Martti Kemppinen. Väsyttävästi kuormitettujen liimaliitosten testimenetelmän kehitys. 2012. 45 s. (nid.)

A:71 Minna Malankin. Venäläiset matkailun asiakkaina. 2012. 114 s. + liitt. 7 s. (nid.) 114 s. + liitt. 7 s. (pdf)

A:72 María del Mar Márquez – Jukka Mäntylä. Metsätalouden laitoksen opetussuunnitelman uudistamisprosessi. 2012. 107 s. + liitt. 17 s. (nid.)

A:73 Marjaana Kivelä (toim.). Yksin hyvä – yhdessä parempi. 2012. 115 s. (nid.) 111 s. (pdf)

A:74 Pekka Hartikainen – Kati Kontinen – Timo Antero Leinonen. Metsätiensuunnitteluopas – metsä- ja piennartiet. 2012. 44 s. + liitt. 20 s. (nid.) 44 s. + liitt. 20 s. (pdf)

A:75 Sami Luste – Hanne Soininen – Tuija Ranta-Korhonen – Sari Seppäläinen – Anni Laitinen – Mari Tervo. Biokaasulaitos osana energiaomavaraista maatilaa. 2012. 68 s. (nid.) 68 s. (pdf)

A:76 Marja-Liisa Kakkonen (toim.). Näkökulmia yrittäjyyteen ja yritysyhteistyötoimintaan. 2012. 113 s. (nid.) 106 s. (pdf)

A:77 Matti Meriläinen – Anu Haapala – Tuija Vänttinen. Opiskelijoiden hyvinvointi ja siihen yhteydessä olevia tekijöitä. Lähtökohtia ja tutkittua tietoa ohjauksen ja pedagogiikan kehittämiseen. 2013. 92 s. (nid.) 92 s. (pdf)

A:78 Jussi Ronkainen – Marika Punamäki (toim.). Nuoret ja syrjäytyminen Itä-Suomessa. 2013. 151 s. (nid.) 151 s. (pdf)

A:79 Anna Kähkönen (toim.). Ulkomaalaiset opiskelijat Etelä-Savon voimavaraksi. Kokemuksia ja esimerkkejä. 2013. 127 s. (nid.) 127 s. (pdf)

A:80 Risto Laukas – Pasi Pakkala. Suomen suurimpien kaupunkien metsätaloustoimintojen kehittäminen. 2013. 55 s. + liitt. 8 s. (nid.)

A:81 Pekka Penttinen – Jussi Ronkainen (toim.). Itä-Suomen nuorisopuntari. Katsaus nuorten hyvinvointiin Itä-Suomen maakunnissa 2010–2012. 2013. 147 s. + liitt. 15 s. (nid.) 147 s. + liitt. 15 s. (nid.)

A:82 Marja-Liisa Kakkonen (ed.). Bridging entrepreneurship education between Russia and Finland. Conference proceedings 2013. 2013. 91 s (nid.) 91 s. (pdf)

A:83 Tero Karttunen - Kari Dufva. The determination of the mode II fatigue threshold with a cast iron ENF specimen. 2013. 24 s. (nid.)

A:84 Outi Pyöriä (toim.). Vesi liikuttaa ja kuntouttaa - hyviä käytänteitä vesiliikuntapalveluissa. Tuloksia VESKU-hankkeesta. 2013. 63 s. (nid.) 63 s. (pdf)

A:85 Laura Hokkanen - Johanna Pirinen - Hanna Kuitunen. Vapaaehtoistyö, kansalaisjärjestöt ja hyvinvointipalvelujen kehittäminen Etelä-Savossa – esiselvitys. 2014. 114 s. (nid) 114 s. (pdf)

A:86 Johanna Hirvonen. Luontolähtöisen toiminnan hyvinvointivaikutukset ja niiden arviointi. Asiakasvaikutusten arviointi Luontohoiva-hankkeessa. 2014. 70 s. (nid.) 70 s. (pdf)

A:87 Pasi Pakkala. Liiketoimintaa ja edunvalvontaa – Näkökulmia työhyvinvointiin metsähoitoyhdistyksissä. 2014. 52 s. (nid.)

A:88 Johanna Arola - Piia Aarniosalo - Hannu Poutiainen - Esa Hannus – Heikki Isotalus. Open-tietojärjestelmä. Etämonitoroinnin kehittäminen osana ympäristöteknologian koulutusta ja innovaatiotoimintaa. 2014. 71 s. (nid.) 71 s. (pdf)

A:89 Tapio Lepistö. Luonnonkuitukomposiitit. 2014. 67 s. (nid.) 67 s. (pdf)

A:90 Kirsti Ilomäki - Kari Dufva - Petri Jetsu. Luonnonkuitulujitettujen muovikomposiittien tutkimus ja opetuksen kehittäminen. 2014. 49 s. (nid.) 49 s. (pdf)

A:91 Jaana Dillström - Erja Ruotsalainen. Huomaan, että osaan. Opiskelijoiden kokemuksia simulaatiosta. 2014. 46 s. (nid.) 46 s. (pdf)

A:92 Kati Kontinen. Huonosti kantavien maiden ja teiden vahvistamisratkaisut. 2014. 39 s. (nid.) 39 s. (pdf)

A:93 Mika Liukkonen - Elina Havia - Henri Montonen - Yrjö Hiltunen. Life-cycle covering traceability and information management for electronic product using RFID. 2014. 55 s. (pdf)

A:94 Liisa Uosukainen (ed.). Open source archive. Towards open and sustainable digital archives. 2014. 90 s. (nid.) 100 s. (pdf)

MAMK

University of Applied Sciences

Publisher: Mikkeli University of Applied Sciences
Publication serie: A Research Reports - Tutkimuksia ja raportteja 94