Towards a National Digital Repository – Case Mikkeli

By Osmo Palonen

The City of Mikkeli is capital of the Province of Eastern Finland. The people living there (close to 50,000), in the heart of the lake district, get their living from services, government, regional and local administration, and some industries like graphic arts are important as well. There are not very many fields where Mikkeli is among the largest in Finland, but the local archive and library community is the second largest in Finland. At the national level operate the Central Archives of Finnish Business Records (ELKA) and the National Library's Microfilming and Conservation Centre which is also the National Digitization Centre. The Provincial Archive of Mikkeli serves the south-eastern region including most of the archive material of the ceded Karelian region. Mikkeli Polytechnic is now growing towards national level status, as the R&D and education centre for digital archiving, preservation and digitization. This group of four supported by the local and regional authorities has the target of becoming the national leader in digital archiving and digitization. In this task all the organisations also co-operate with their national partners.

Mikkeli Polytechnic as the leading higher education operation in Southern Savo region is serving the community. Digitization, digital content management and digital archiving is one of the key fields in the strategy of Mikkeli Polytechnic. Since 1999 we have spent close to six million euro on projects to create the foundation for digital archiving and preservation. Most of the funding has come from the European Union’s Objective 1 programme (ESF and ERDF) that includes national funding from the Finnish Government via the Province of Eastern Finland. In addition we have used our own and local funding as well. There have been 10 to 15 persons working for these developments on average.

In the long run an education and research organisation like ourselves will not be the prime mover of the national digital repository. A company or foundation, whose majority is owned by the public sector, will most probably take the reins of the operation within a few years. Our role is to make the development and start the operation. After that our role is in education and R&D, especially in information technology for digitization and preservation.

Born digital – archiving on paper

Digital archiving does not have a long tradition and it is not obvious how to manage digital material. Over 90 percent of the material is created today by using digital tools i.e. word processing, CAD, e-mail and spreadsheet software, for archiving it is printed on paper. These papers are held in the archives inside the organisations. If we would like to copy these procedures direct to the digital world there has to be a digital repository in every organisation.

I will give an example: In Finland we have 430 cities, towns or other municipal organisations, and in addition there are over 200 other organisations owned by those 430. Should we have 650 digital repositories to hold the public records of local authorities? 650 RAID-boxes or storage systems and 650 back-up and archiving processes? If we are to keep those records available in the future we should have a migration strategy in the 650 organisations and use converting software for TIFF's, PDF's and other contents. IT-companies would love it, but the tax-payers would not. And this was only for the self-governing municipalities.

In 2003, when I started at Mikkeli Polytechnic, there was a discussion in Finland whether the digital contents should be archived on a provincial basis or by the sectors. There are six provinces in Finland, and if using a regional model would have to establish six digital repositories; if using the twelve ministries or offices it could have been twelve digital repositories, again with 12 storage systems back-up processes and 12 separate migration and converting

operations. Compared to 650 this model was far better, but still we only have five million Finns. Do they have 11 complete digital repositories in the UK? I would say one is fine. In Mikkeli we see our current activities as the predecessor of this one and only digital repository.

National archives and the network of provincial archives are still important to provide citizens and organisations close access to material held in the government archives. It takes time before the most interesting material is in digital format. It is the same with the municipal, private and business archives. In the digital age we have a couple of new add-ons to use. We can distribute digital information via communication networks close to every home and office. We can now better manage the new types of records which can hardly be printed. For decades history has not only been written, it is created as photographs, sound, moving images and in the last wave multi-media contents and web-sites. Using digital preservation we can preserve that information as long as traditional records.

I learned in my previous life, serving the newspaper industry, that when adopting new information systems there has to be a considerable effect on the operational process, or you waste your money. This is also the case in digital archiving and preservation. I am still really not saying that all what we have learned should be forgotten, vice versa. The archivist has, over the centuries, created excellent methods and archiving theory for records' management and preservation. The first migration process is to convert those cornerstones to the foundation of digital preservation. This is what we are trying to do in our tiny operation in the wider archiving world. In the following chapters I will try to relate what kind of rules we have created and how well we have been able to follow those. In the last part of this case study I speak about the projects and services we are now running, you can evaluate if we have misinterpreted the tradition.

The repository plan

Mikkeli Polytechnic has laid the first foundations to create a digital repository to be available for the nation. We already held the records and documents from the health sector, municipalities and business archives. Up to now most activity has been in projects funded by the EU, but in the near future the priority will be in contracts with the public and private sector. There will always be projects to develop new methods and tools, but the bread and butter should come from the real world.

The principal ideas we have used in building the digital archiving processes and the repository are as follows. The essential rule has been that we cannot rely on any single supplier, as we have to keep the contents available. We also decided to prefer open standards. In practice to follow this is sometimes difficult, and we have not completely succeeded in this, as you can find out.

Our model of operation can be split into four levels:

1. Off-line level, media

At the lowest level of preservation we now use LTO-2 tapes as the archive copies (at least two) in which all the document originals are saved, not dependent on those being radiological pictures, digitized videos or official records from municipalities. The life cycle of LTO-2 is expected to be five years; after that the documents will be moved to the widely supported format of that time, maybe LTO-4 or LTO-5. During these five years, we need to check at least once that the tapes are fully functional. We also may consider using IBM WORM-tapes and standard data DVD's as customer copies. The material that will only be saved on these off-line archive copies, is the archive copies of moving images. Using high quality we can get 180 gigabytes per hour, i.e. in one of our projects more than 100 terabytes.

2. SAN-storage

The second level in digital repository is the redundant disk storage system and the on-line tape library based on the SAN network. It is planned to have most of the archive material near-on-line and distribution copies either on disk or tape. There are two copies of material in near-on-line library tapes and for hard disks we use RAID-5 mirroring. There is a standard back-up from the disks in the tape library. The next step we are currently planning is to make the SAN storage systems, including disks and the tape library redundant. We are looking for a location inside the Fennoscandian Shield some kilometres away. When this step is complete we will have the information in two off-line copies, two in mirrored disk systems using both RAID-5 mirroring and two near-on-line tape libraries. With the software we can select how to partition the space.

3. Servers and operating systems

Servers for databases and the critical applications are clustered. We decided to use Windows Server 2000 and 2003 and Linux operating systems. The database engine used is the MS SQL Server 2000, but we are considering using Open Source relational databases like PostgreSQL. The user recognition and authentication is based on LDAP and Novell products. When we make the step to two in SAN storage, we will also split the server clusters, one server in the first and one in the second location, this makes the services even more reliable.

4. Applications and formats

All the metadata is in two separate systems in XML and we require that it can be opened by archive software or even read by a browser or a text editor. Text documents are saved either as XML with style sheet or just plain text. The look and feel of the documents including text and pictures of graphics is saved as PDF (as PDF/A when available). The metadata contents and structured documents are transferable and applications are based on open standards. Metadata and data files can be certified by organisational digital signature or by server signature.

The audiovisual contents are saved in non-proprietary formats so that those are codec-independent. We only use IT-based media. The digital archive of sound files is based on broadcast wave and MP3 format, and moving images on MPEG-2 and MPEG-1. Photos are saved in JPEG or TIFF and plans and drawings in PDF or TIFF; JPEG2000 is assumed as the next step.

The radiological pictures with metadata are in DICOM format, pictures in TIFF, compressed 1:2 or 1:3. The most important thing here is the reliability. The system cannot be down and we were looking for 99.99 uptime. We are prepared to receive medical records in HL-7 messages using CDA Release 2 format. These contents can be managed via the hierarchy used in X-Archive.

… and realisation

When we were looking for the solutions to our repository one requirement was that the companies involved were really committed to digital archiving and preservation. On levels 1 and 2 we selected IBM DS4500 hardware with 3584 tape library and Tivoli data management software. We now have 11.5 TB disk and 50 TB tape space. The disks are both FC/SCSI (8 TB) and Serial ATA (3.5 TB). The limits for this single system are 32-56 terabytes on disk and 400 terabytes on LTO-2 tape. The storage system and the first servers were installed at the end of 2003. The server hardware is also from IBM.

Making decisions on the level 3 was most difficult. In the end we selected two Finnish software companies and required that their products should be integrated. AvainTechnologies had XML-based tools for archivists to manage the permanent and long term material (X-Archive). They

also had experience in XML-forms, digital signature and PKI authentication tools were already in use and tested, but their metadata ingest and search functions were either completely missing or not at a very high level. Another company called Profium had an excellent RDF-based information management engine called SIR (Semantic Information Router) already in use in some archive applications and news agencies. Profium had also been involved in consulting work for the Sähke-specification.

The workflow starts with archive material (other than the PACS system) from the automatic ingest of the contents from the watched folders via SIR input handlers, these can now handle broadcast wave and XMP used in Adobe-based files. Manual upload is done and the metadata is completed by user interfaces to SIR to build by ourselves. The contents are saved in the storage system and the metadata is copied via SOAP to X-archive. A search can be done by RDF-based user interfaces utilizing SIR functionality and the archivists can use X-archive, who has the archive hierarchy and rules to manage the preservation of the contents. SIR will also be used as the automatic converter. Looking back now, 14 months after the system start, we underestimated the work needed to build input handlers and user interfaces. Other than that the system is what we expected, even more. Regarding X-archive we are just getting the next generation with the improved functionality and then that part of the system can also get fully used.

For the PACS system we required high availability and reliability to keep the maintenance costs low and service quality high. The radiological archive is Sectra from a Swedish company. The same software manages picture distribution and radiologist workstation services; that is we provide a complete PACS service. The examinations in Southern Savo health district are produced in the hospitals or health centres and sent via LAN and a dedicated Ethernet connection (1 Gigabit/s) to our servers. In the national contract we use the same Sectra system to serve a mammography screening process. The job is done by a company owned by the municipalities. The examinations are now done in three permanent screening centres and there will be seven of those in total by the beginning of 2006. There are now also two mobile units and before the end of 2006 there will be more than twenty. The permanent centres are connected to Mikkeli via a 100 mbit/s network and the mobile units are using ADSL connections.

General store?

At the moment we have different types of archive material and we hope there will be even more of those. To manage different sorts of material in different formats and from various branches is one of our principal ideas. The reason is simple: the more bytes, the cheaper per byte. Even more important is, that the migration of the contents in the digital archive has to be under one common control. We also can see the benefit of understanding different types of materials, applications and customers. In healthcare the privacy and data security are numbers 1 and 2, in business content we can use the same methods and practices. But we can also bring the technical metadata knowledge learnt in the work with audio files, when creating photoarchive for a hospital. I have heard many doubts about this multi-field approach from those who look at the world from their profession only and I'm glad to say that I have seen that the archivists has proved to have a broader view. This summer I studied the historiography of science, where I have seen the same kind of fences that are built in the minds of people: some think that the scientist is the only one able to write the history of science, but some others think that only a professional historian should be allowed and able to do that. I didn't agree with either of those doctrines.

Our launch might have been too early – or too late depending on the viewpoint. The problem is that the methods, practices and tools for electronic records' management are still under way. You have already heard from Mr. Markus Merenmies that the National Archives of Finland have

just published the rules to archive digital contents inside the Government organisation. According to that, it will take several years before those records' management systems are transferring archiving contents to National Archives. That does not mean that we have to wait. There is much material in other fields waiting for proper preservation.

ELKA goes digital

The first special project which we started in the beginning of 2004 is called ElkaD. This is a joint project with the Central Archives for Finnish Business Records (ELKA). Located in Mikkeli ELKA has been a national repository for business records for 25 years. There are 18 kilometres of paper archives including the essential archive material from the Finnish enterprises in industry, trade and services. The business records are not only paper documents, in 2003 there was about 1200 hours of audiovisual material on magnetic tapes and films, a lot of mechanical, electrical and other drawings from ship and coachbuilders and industrial buildings. With this so called special material the biggest worry was how to preserve analogue sound and video recordings on magnetic tapes. The majority of that material was already under threat of losing the information. Based on our experience the colours have already become faded and the sound hazy. If another decade is lost there will be not much to preserve.

However, the job was not only done for this project. As a result we test and keep documentation of the methods and best practices to save this kind of material in digital format and keep it available for researchers and the public. Part of the material was still under a limited right of use and that has also created some lessons for us to learn.

Until now we have found that digitization of sound is quite a straight forward process, but video is continuously creating problems for us. Those problems might arise from our decision not to use high-end solutions. We use Pinnacle Crome and Liquid Edition workstations. Chrome is working quite well with professional tapes like U-Matic and our 8 mm film scanner, but with standard VHS the number of vertical stripes did not fit. We held up our hands and started to use Liquid instead. There too some of the material is running well, but sometimes it drops frames and the sound can be out of sync. It's no wonder that the digitization of video in non-proprietary formats has not been very popular. Still we think that we have succeeded in using the MPEG-2 and MPEG-1 formats.

Digitization of business operations has created another big threat to historians and archivists. No letters are written and e-mails disappear. In ELKA there are continuous collections of records from companies in analogue format, but in the 1990s those have not been printed on paper, but kept in databases only. If the information is now in databases, the chain will often be broken. To print databases on paper only creates benefits for paper manufacturers. Information that was left in obsolete systems can hardly ever be used unless the researcher can and be willing to read thousands of printed database tables or sequential flat files and system descriptions.

Our first database to be preserved is the Musa database that holds information on performers and the copyrights of all radio programmes broadcasted by the Finnish Broadcasting Company (YLE). Between 1991 and 2004 this information was saved in a DMS-based MUSA-database and will now be archived in structured document format. Also all the system and project information will be digitized. Because information in Musa is still in needed i.e. when the programmes are repeated, the new normalised archival database using ISO-SQL will be created. This database will be open for research and internal use in YLE the same way as the earlier music reports on paper and now on the shelves in ELKA. The difference to earlier times is that the information is available via a network. In born digital material YLE has another partnership with the Elkad project. There will be minutes from internal executive teams to be saved in Mikkeli Polytechnic. Most of those have also been archived on paper in ELKA.

Digital photographs are now widely used in industrial companies. Instead of using professional photographers a lot of documentation is done by using digital cameras and the file servers are often full of picture files. Even when only a small share of those will be kept in archives, this new format has created problems for traditional photo archives. The digital photos – as well as all other digital formats – are carrying with them valuable technical metadata that will be needed in the conversions to the formats used in the future. To save this metadata and to evaluate the needs for the photo archives in the future Mikkeli Polytechnic and ELKA has made a partnership with UPM Kymmene, a Finnish forest industry giant. In this part of the Elkad project we test the new photo archive application. The leading idea in that is that all metadata should be input into the JPEG files using Photoshop XMP and transferred automatically into the RDF-structure of the archiving software. We'll see how well the photographers and people in marketing and documentation can and will key in the basic metadata! The fields described in this application are:

- Headline
- Caption
- Location
- Taken by
- Temporal (start and end by date)
- Controlled vocabulary
- Category defined by the customer

To be honest, there is a user interface for archivists to add and edit the metadata.

Municipal digital archive

As I already said there are 430 self-governing urban or rural districts, who are in charge of their own record offices. Those municipal organisations can have from 131 to 559,000 members and also the organisations can be very different. In the paper age the difference between the small and the large has been in the volume of archiving material – and the number of archivists, of course. In the digital age the dissimilarity is much larger: the big cities can, if they prefer, build a complete digital repository, but the small ones just cannot afford a proper digital archiving system. A file server with a DAT-backup tapes is not a digital repository.

In the project Kunda we will create together with the local government of Pieksämaa, a municipality of 8.700 inhabitants, a model of how the small and mid-sized municipalities can move into the digital age. The results of the project are planned to be the basis for the national archiving services for municipal organisations. The target is in the future, but we also look back. For the future we work to ingest the born digital material, based on the results of the Sähke project and the study made in the project concerning the plans and drawings. Looking back we digitize the material created by the municipal administration boards and councils. There will be approximately 50 shelf metres of material to be digitized. Pieksämaa is a new municipality that was created in the beginning of 2003 by the joining of three smaller rural districts. One of those old archives will be digitized as far as reasonable, in one we will predict which is the most used material and digitize it and in the last we will not touch it. When we know the cost of digitization we will measure the usage of the digitized and un-digitized material and see if that can be utilised better than the material still only on paper.

ASP-Services for health care

The first paid contracts were done in the health sector. We started as the Application Service Provider for the radiological picture archive for one regional and one national organisation. The preservation time for these pictures is only ten to twenty years, but there has been a request to

save 6.6 percent permanently for research purposes. The regional contract is based on 75,000 and the national one close to 200,000 examinations per year. In terabytes this is over 40 per year. Archiving of the material is only one part of these contracts, the other part is to provide picture distribution for clinical users and special workstations for the radiologists.

We broke the general rule of our own and the archiving community with these contracts. We know that information systems in active use cannot be used as archive. We have some explanations: we estimate that the PACS system with upgrades can be running for more than 20 years and we are sure that within a few years we will have solved this problem. Also our responsibility described in the contracts is for five years at a time as a maximum; if the contract terminates, we just provide the material to the customer in digital format by using the DICOM standard on tapes. However, to fit in with our migration path we are planning to develop a metadata-based archive solution for pictures created in health care organisations by using CDA R2 and DICOM information.

When starting a digital repository it is important to get customers that create the critical mass of terabytes and pay for the services from the beginning. In health care 80 percent of the data to be saved in archives is coming from radiology. This has created the need for a new approach especially when the IT directors of health districts are not willing to keep the material for 20 years when it is used very seldom.

In health care we will also continue to archive the complete medical records. Data created in those systems have a preservation period close to 100 years and it is clear that the systems in active use cannot be used as permanent archive. The role of archiving can also be different here. Archive can also act as an integrator of the systems. There could be hundreds of systems used in provincial health care organisations. Digital archive can receive and distribute information between the systems much easier than if those hundreds of systems are communicating between each other. The national health strategy in Finland assumes that the digital medical records can be shared even when the patient is moving or travelling to the other provinces. To achieve this, the systems should fulfil the strict requirements of privacy and security in data management.

Enhance the role of archives

To summarise our experience at the moment, I can see large benefits in digital archiving. The best thing is that we can provide the material from the repositories to the researchers own computers as well as to the public. In this role the archiving community can be one interesting source for the digital contents requested by the information society and part of the democratic process in where the citizens – and also the administration - can easily get accurate information on their interests.

I am also sure that the concept to centralise the preservation services in fewer digital repositories and use those resources decentralised is the way to go. The information highways between the countries will soon allow movement of the contents, the national repositories can be a distributed network like those the Distarnet or LOCSS projects have predicted.

In this European forum I would like to make a forecast that the electronic archives of the European Union could be shared in digital repositories in three to five locations. We will try very hard to ensure that one of those will be in Mikkeli.