MANAGING RISK

REPORT

THE NATIONAL LIBRARY OF NORWAY

LONGREC CASE STUDY: REPOSITORY RECORDS MANAGEMENT

DNV REPORT NO 2008-0273

DET NORSKE VERITAS

REPORT

Date of first issue:	Project No:	DET NORSKE VERITAS AS Research and Innovation
2007-07-05	91303021	
Approved by:	Organisational unit:	C3 1322 Høvik Norway Tel: +47 67579522 Fax: http://www.dnv.com NO 945 748 931 MVA
Inger-Mette Gustavsen	Research and Innovation	
Author:	Client ref.:	
Olga Cerrato, Lars Gaustad (NB), Jon Ølnes	The National Library of Norway (NB)	

Summary:

This case study is a part of the LongRec (2006-2009) research project, http://research.dnv.com/longrec. This report is the first delivery of the case study. The case study addresses the challenges faced by the National Library of Norway regarding content management, primarily migration and conversion, of the National Library's trusted digital repository. The main research question is how to be certain that the repository records remain unaltered, i.e. 'survive', despite multiple transitions between hardware and software which are inevitable in the future. The complicating factor for this case is the huge volume and heterogeneity of digital content the National Library operates with, which includes not only written material but also images, various audio files, films and Internet publications. Due to volatility of storage media and technology and due to the novice technology that appears all information objects are migrated to new storage every 3-5 years. This implies a massive copy operation, which takes several months to complete at present.

The writings of this report were based on interviews with the case partner and documents submitted by the case partner. The report describes the present state with respect to the repository records management, identifies issues that may need improvement and then narrows down to a concrete study topic that LongRec and the National Library will concentrate on further in the project.

DNV Report No:	Subject Group:	Date of this revision:	Revision No:	Number of pages:
2008-0273		2008-02-15	4	20

Report title:

The National Library of Norway, LongRec Case Study: Repository Records Management

REPORT

Table of Contents *Page*

REPORT

1 INTRODUCTION

1.1 The LongRec project

This case study is a part of the LongRec (Long-Term Records Management) project run by Det Norske Veritas (DNV) in collaboration with a number of case partners, commercialization partners and research partners. The primary objective of LongRec is the *persistent, reliable and trustworthy long-term archival of digital information records with emphasis on availability and use of the information*. The project's public web site is at http://research.dnv.com/longrec/

LongRec is a three year project (2007-2009) partly funded by the Norwegian Research Council. The project constitutes the Norwegian team of the InterPARES 3 project, http://www.interpares.org

LongRec addresses several research challenges[1], each of which is assigned a short name (in parentheses below): records transition survival (READ), long-term usage (FIND), preservation of semantic value (UNDERSTAND), preservation of evidential value (TRUST) and legal, social, and cultural framework (COMPLIANCE). Each research challenge is addressed by:

- General studies compiling state of the art and best practice of the area.
- Research on selected sub-topics, performed by the research partners and by one PhD student for each research challenge.
- One or more case studies with LongRec case partner(s).
- Studies on opportunities for products and services at commercialization partners.

1.2 The case study

This case study addresses the READ (records transition survival) research challenge by investigating the challenges faced by the ***National Library of Norway (NB)*** regarding content management (primarily migration and conversion) of NB's repository. Each information object in the repository is stored in three copies; at present one on-line on disk and two on tape. The number of objects and the amount of information is rapidly increasing. Some objects are very large. NB expects to have close to two petabytes of data by the end of 2008.

Due to volatility of storage media and technology, and to be able to utilize the latest in technology, all information objects are migrated to new storage every 3-5 years. This implies a massive copy operation, which at present takes several months to complete. Challenges that are relevant to this scenario are:

- Are there measures that should be taken to increase the safety of the migration process? Even though the risk of loss or corruption of an object is small, the sheer volume of the migration still leaves some uncertainty. Information loss is in principle unacceptable.
- What are the implications of introducing further content management operations, presumably integrated with the migration process? In particular, sooner or later formats of information objects will become obsolete, raising requirements for conversion of all objects of given formats. Maintenance of metadata may also be desired, including metadata formats.
- Are there limits to the feasibility of such a migration process? Considering bandwidth for media access, processing speed of computers, lifetime of media, and other parameters, at

[1] We refer to the project's web site http://research.dnv.com/longrec for a description of the research challenges.

approximately what stage (information size and number of objects) does the current migration scheme become practically impossible?

These challenges are further outlined in the following. The third point is not specifically addressed by the case study. The purpose of this first case study report is:

- Identify the present state at the case partner and identify issues and topics that may need improvement.
- From the identified issues and topics, narrow down to a concrete case study topic. Describe desired situation as far as possible for this topic and discuss ideas and other input pointing at directions for solutions.

Then, a second case study report will give recommendations. A decision will also be taken regarding piloting (practical implementation and testing) of the recommendations.

NB's expectation to the final case study results is: A document containing a description and a verification of best practises in repository infrastructure maintenance and format preservation.

1.3 Audience and accessibility

The LongRec case studies serve multiple purposes:

- Most important, the case study partner must benefit from the results.
- Then, the work should be of value to the partners of LongRec and to the general research carried out by the project.
- Preferably, the results should also be available and of interest for other parties, such as partners of the InterPARES project.

This report is publicly available and contains sufficient background information for a general audience with a reasonably good background in the area.

2 DESCRIPTION OF THE CASE PARTNER

The National Library of Norway (NB) is the nation's memory as well as a multimedia information centre. NB preserves and distributes the nation's heritage as it exists in handwritten works, maps, books, periodicals, newspapers, photographs, films, broadcasting, music and internet publications. NB's goals are:

- Be among Europe's most exciting and modern national libraries.
- Form the core of the Norwegian Digital Library.
- Offer high quality knowledge and experiences.
- Assist in the understanding of culture and technology.
- Be an organization willing and able to change.

NB has some 340 employees in Oslo and Mo i Rana. The 2007 annual budget was approximately 280 million NOK (about 35 million €). About 1/10 of the budget is used for investments. For more information see: http://www.nb.no/english/annual_report

NB is to preserve and make accessible to the present and the future the information that shapes the Norwegian society, regardless of how and in which medium it was published. A main pillar in the

REPORT

collection of materials is the *Legal Deposit Act*. According to this act, specimens of all information published in Norway shall be handed over to NB.

- The first act appeared in 1697
- The present act came into force in 1990.
- The act covers all types of media, including digital documents and broadcasting.

The present act gives NB a broader scope than most other national libraries in that not only printed material shall be handled. All information produced for public availability is covered by the act regardless of the original medium, so the collection of the NB covers everything from printed material, music, broadcasts, films and the web. (Note that on a national level, the National Archival Services of Norway are responsible for archive material, while NB handles published information.)

In addition to material received according to the act, NB purchases or otherwise receives historical material, in part to make its collections complete, in part to maintain lending collections. NB owns and manages several unique collections. All are available for research and documentation, and most are accessible to the public through NB's general library services or via the Internet. These include:

- unique manuscript collections (including handwritten manuscripts),
- special book collections,
- music collections,
- radio broadcasts from the 1930s up to the present day,
- film collections,
- theatre collections,
- a large map collection,
- posters,
- photographs,
- newspapers.

NB has embarked upon the process of digitizing ALL of its collection for preservation and access purposes. How copyrighted material will be digitized and how access will be granted will be decided in a dialogue with the rights holders.

The Norwegian top level domain (.no) of the World Wide Web is regularly harvested and archived in NB's repository. In 2007, approximately one billion web pages were downloaded.

NB shall receive and store all material broadcasted by the state owned National Broadcasting Corporation NRK[2]. NRK's radio archives are being digitized and stored at NB. Digitizing of TV broadcasts is not yet prioritized for capacity reasons. NB is entitled to receive all broadcasted material from other sources than NRK and has decided to demand this from actors that cover all of Norway, but is requesting only samples of material from local broadcasting actors.

[2] Norwegian: Norsk Rikskringkasting, http://www.nrk.no

REPORT

3 GLOSSARY OF TERMS

Archival Information Package (AIP) – is an Information Package, consisting of the Content Information and the associated Preservation Description Information, which is preserved within an archival information system (OAIS def.).

Authenticity – the trustworthiness of a record as a record; i.e., the quality of a record that is what it purports to be and that is free from tampering or corruption (InterPARES glossary).

Conversion – the process of changing something from one form or medium to another, while leaving the intellectual content unchanged (InterPARES glossary).

Dissemination Information Package (DIP) – is an Information Package, derived from one or more AIPs, received by the Consumer in response to a request to the archival information system (OAIS def.).

DSM– Norwegian: Digitalt Sikringsmagasin – a software package developed internally in NB that handles the safe storage of NB's digital objects.

Information Package – is the Content Information and associated Preservation Description Information which is needed to aid in the preservation of the Content Information. The Information Package has associated Packaging Information used to delimit and identify its components.

Ingest – is the OAIS term used to describe services and functions that accept Submission Information Packages from Producers, prepare Archival Information Packages for storage, and ensure that Archival Information Packages and their supporting Descriptive Information become established (OAIS def.).

Integrity – the quality of being complete and unaltered in all essential respects (InterPARES glossary).

Metadata – information that characterizes another information resource, especially for the purposes of documenting, describing, preserving or managing the resource (InterPARES glossary).

Migration of records – the process of moving records from one system or storage medium to another to ensure their continued accessibility as the system or medium becomes obsolete or degrades over time (InterPARES glossary).

OCR – Optical Character Recognition – computer software designed to translate images of typewritten text into machine-editable text.

Refresh – to convert storage of digital components from one medium to another or otherwise ensure that the storage medium remains sound (InterPARES glossary).

Storage Area Network (SAN) – is an architecture where remote computer storage devices (e.g. tape libraries, disk arrays, optical disk libraries) are attached to servers in such a way that, to the operating system, the devices appear as locally attached.

Submission Information Package (SIP) – is an Information Package that is delivered by the Producer to the archival information system for use in the construction of one or more AIPs (OAIS def.).

Validation of file format – verifying that the file is in compliance with the specifications of its purported format.

REPORT

4 NB'S DIGITAL REPOSITORY – STATUS AND CHALLENGES

4.1 Architecture

The application maintaining NB's repository is called DSM (Digitalt SikringsMagasin), a set of software tools developed internally to securely store digital objects, in InterPARES terms known as records. These applications provide among other things each object with a unique identifier and attach preservation metadata that is believed to be needed in the maintenance of the content and for rendering it in the future in ways that preserve authenticity. An MD5 checksum is created and stored with the object. The replaceable hardware system in use stores three copies of each object on two different technologies in two different localities. A new system (2007) using a RAID5 disk array is used for one set of the digital objects, while the other two instances are kept on tape robots. Integrity in restore operations is based on the operating systems. The switch to the new system also included a change to the SAM-FS file system.

NB's architecture and implementation are according to the OAIS reference model[3] (ISO 14721:2003); this is shown in Figure 1 below.

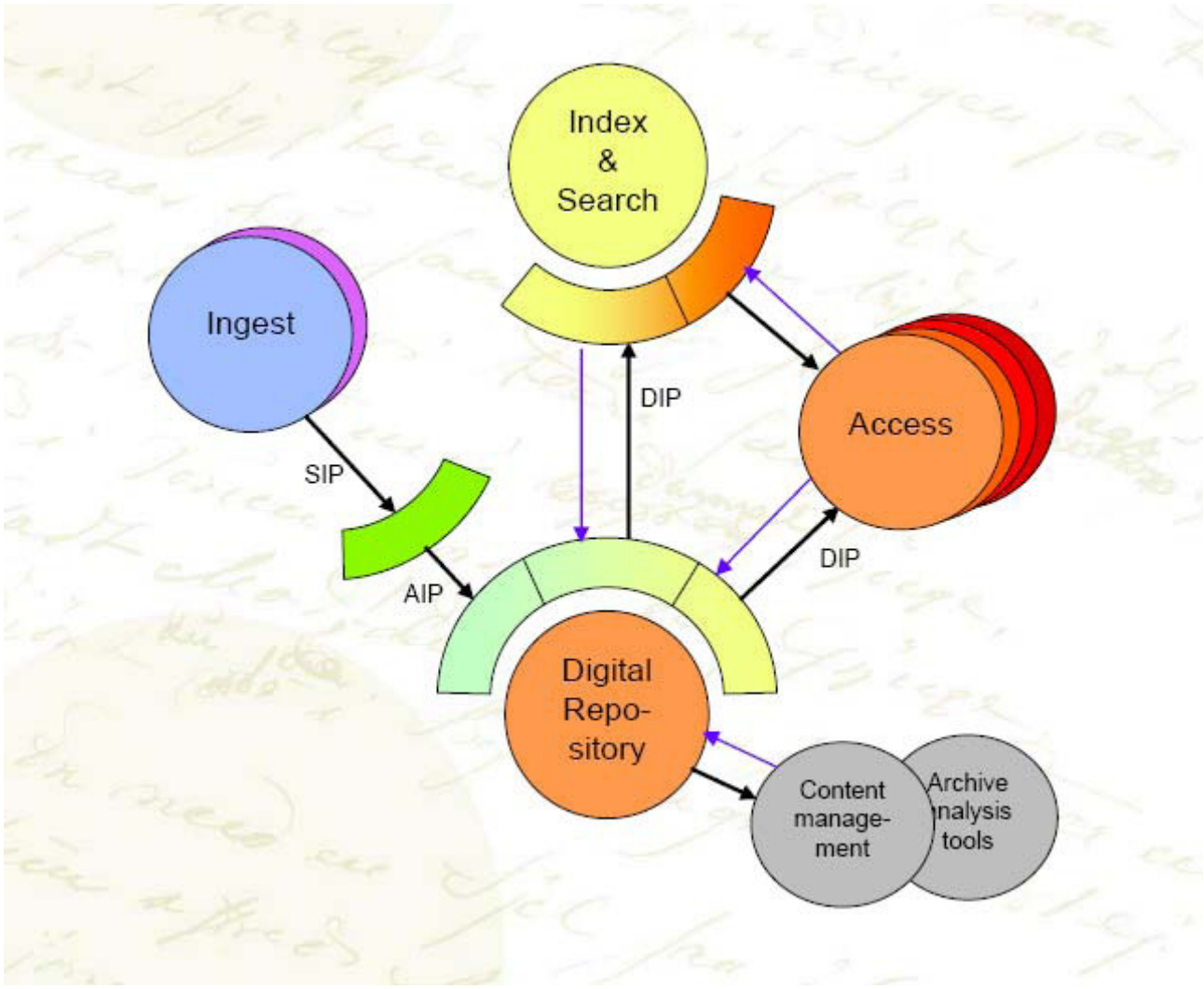

Figure 1: NB's digital repository model

[3] http://public.ccsds.org/publications/archive/650x0b1.pdf

REPORT

As shown the model references three different packages:

- SIP - Submission Information Package
- AIP – Archival Information Package
- DIP – Dissemination Information Package

The OAIS functional model defines six areas of concern:

- Ingest
- Data Management
- Archival Storage
- Administration
- Preservation Planning
- Access

The relations between the areas are shown in Figure 2 below.

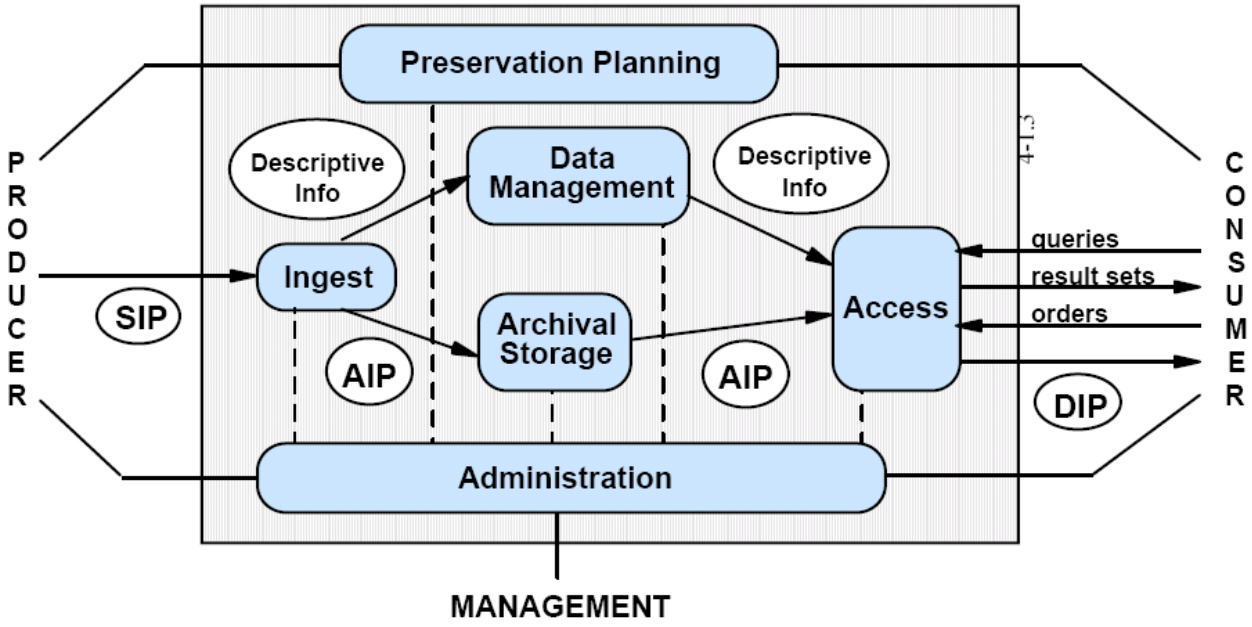

Figure 2: Areas of concern in the OAIS functional model, and their relationships.

NB needs to ensure the best possible transparency with other repositories. This means state of the art procedures for ingest and storage that comply with international standards and recommendations; in other words, a solution that will comply with international initiatives on the certification of trusted digital repositories. (See for example the Certification of Digital Archives Project http://www.crl.edu.) This requirement goes for the whole process, including migration, conversion and other content management; the whole process needs to be trusted (authenticity, integrity, security).

NB is in a fairly good position with respect to this goal but is not certified. Recommendations and standards are continuously evolving, which means that NB must follow, or preferably even actively participate in, relevant work on the international scene.

4.2 Ingest

Building on the OAIS model AIPs are created with preservation metadata attached though use of extraction tools like JHOVE[4] and DROID. Data integrity is monitored on file movement through fixity checks, all processing to the file is registered as events in the PREMIS[5] schema and all relevant preservation metadata are also imported into NB's bibliographic catalogues. These catalogues are outside of DSM[6], and NB also refers to completely external catalogues like BIBSYS.

There is no integrity checking in the transfer of files from the production stage to the storage area, whether the files are produced in-house or externally, except for in-house digitalization of photos.

Some examples of ingest processes are:

- Books are digitized at NB to preservation resolution TIFF format, which after OCR (Optical Character Recognition) and structural analysis (identifying chapter headings and the like) are migrated from TIFF to j2k (lossless compression) without integrity checking apart from initial testing (convert samples from j2k back to TIFF and compare to the original TIFF version).
- A lot of legal historical radio broadcast material is stored on QIC (Quarter Inch Cartridge) tapes, which by now is an obsolete system, in the proprietary MUSICAM sound format 112 kbs. This is converted to MPEG1levelII 192 kbs with sound file length verification only, and these sound files are stored in DSM.
- Legal deposit radio is received from NRK as MPEG1levelII 384 kbs files representing one hour of broadcast, together with an XML export of NRK's production database covering one day of descriptive metadata.
- From the National Library for the Blind and Visually Impaired NB receives files wrapped in the sound book format DAISY[7], one package containing MP3 + SMIL[8] and one WAV + SMIL, with a limited amount of metadata; filename relates to an external database.
- Two major newspapers submit PDF-files with no metadata; the filenames relate to date and page numbering.

The Legal Deposit Act implies that in principle all media and format types have to be accepted by NB. Requirements may be imposed on some sources of information but not on all of them; the same goes for supply of metadata with the information. The resulting situation may be limited availability/accessibility of the information to users. Objects from several producers/owners are stored outside of DSM because the required work to place them inside has not yet been done. The services that deliver these objects to its users are unable to communicate with the DSM so far.

Due to the large amount of objects that need to be processed in most ingest processes NB must in general rely on automated tools for verification, validation and metadata harvesting. Manual procedures are infeasible except for samples and possibly for some ingest processes with a fairly low

[4] Refer to the bibliography chapter for a description of tools, standards and specifications.
[5] Refer to the bibliography chapter for a description of tools, standards and specifications.
[6] An example is the internal system MAVIS (Merged Audio Visual Information System).
[7] See http://www.daisy.org
[8] Synchronized Multimedia Integration Language specification, W3C Recommendation 07, 2001.

volume. The correctness of the ingest processes is crucial – NB needs to be certain that what they store is what they believe they store. This implies trust in the tools used. The different ingest processes and the different input data objects yield different levels of preservation metadata for the storage in DSM.

Although the original data objects (digital, paper or other media) are kept, one cannot be certain that these can be read at a later time. Fixing a broken object in DSM (e.g. detected when someone tries to access the object) by another ingest of the material may not be possible. Sufficient trust in the ingest process could imply that the original object could be disposed of, although the risk would have to be evaluated for each concrete process, information type and medium.

From the present situation of many different ingest processes, NB would like to seek a harmonization and also a more close alignment with international standards and recommendations for trusted digital repositories. Tools like JHOVE, DROID and the NZ metadata extraction tool can be used to create the necessary preservation metadata to populate a set of SIPs/AIPs for each format and producer.

Web-harvesting of the .no domain poses some particular problems. A question is the speed of the current process, which runs twice a year. Some sites should possibly be downloaded more often than others, and arrangements for submission of content (e.g. authors publishing on Internet) instead of scheduled downloading may be investigated. There is insufficient control over formats for web-harvesting. Intellectual property rights and privacy issues block general, on-line access to the archived web content. Norwegian sites under other domains are not covered; some of these may be recognized as being in Norwegian language, and co-operation with other web-harvesting bodies is being investigated. Targeted download processes may also be run in conjunction with particular events, e.g. elections, harvesting relevant sites quite frequently for a limited period of time.

Cross-media publishing is another upcoming, demanding area, e.g. a film, TV-serial, Internet site, computer game etc. that belong together and interact with each other. See below for an idea on use of metadata to organize such composite objects.

4.3 Metadata

At ingest AIPs are created with preservation metadata attached though use of extraction tools like JHOVE[9] and DROID, and static XML-files to populate a METS conforming schema, built on the APSR/NLA[10] METS schema which is built to implement PREMIS schemas and other defined METS schemas like MODS, MIX, LoCs AMD.xsd and VIDEOMD.xsd. All processing to the file is registered as events in the PREMIS schema and all relevant preservation metadata are also imported into the MAVIS[11] catalogue (which is outside the DSM). The AIPs stored in the DSM consist of the information object and an XML file containing the MD connected to but stored independently from the information object.

Metadata harvesting must in general be automated; manual processes are infeasible. The different ingest processes and the different input data objects yield different levels of preservation metadata for the storage in DSM; this depends both on the metadata (if any) supplied with the object and the metadata that can be extracted from the object. NB can set requirements for supply of metadata for some sources but not for all.

[9] Refer to the bibliography chapter for a description of tools, standards and specifications.
[10] METS profile developed by the National Library of Australia (NLA) through the Australian Partnership for Sustainable Repositories (APSR).
[11] MAVIS stands for Merged Audio Visual Information System and is a proprietary database system developed by Wizard Information Services and ScreenSound Australia. MAVIS is being used as a cataloguing solution (an inventory) for a digital collection, e.g. newspapers, audio files, etc.

REPORT

There are several open source tools available for ingest and metadata extraction, with JHOVE, DROID and NZ in use at NB today. JHOVE supports 10-12 formats and NZ 5-6. The most important reasons for choosing the present tools are that they support many formats and that they output a uniform XML format that can be managed for preservation purposes.

Identifications of what has happened to a file are to be found in MD, e.g. what formats the file has previously had. This is especially important for compression formats – e.g. one compression method can be per picture while another can be per group/stream of pictures (called GOP – group of pictures). It is easy to loose information during conversion between such mechanisms.

NB has specified its *Core MD* that in the context of the DSM exists for all digital objects. It is simply ***a minimum set of technical metadata that is required to administer the objects in DSM***. Additional minimum preservation metadata is defined for some formats (TIFF, MPEG, MP3).

PREMIS can add external links in MD – called *relations* – that indicate how *files* belong together. E.g. NB digitalized 24-track studio sound, i.e. this makes up 24 files that belong together. These are stored as one object in DSM. It is possible to refer to other *objects* in DSM as well. An idea is to use such referral to organize cross-media-interactive design and publishing as mentioned above.

Authorizations and access restrictions can now be represented in METS; however this is not in use at NB today. Some materials are open to public access, while others are partly or completely blocked. Scenarios where parts of an object are available and others blocked (an author delivers his book but the photographer does not wish to give up rights on the photographs in the book) should be handled. In addition to intellectual property rights, privacy and personal information protection must be ensured. Today, probably more material than necessary is blocked due to the risk of violation of privacy or intellectual property rights.

Note that a problem with PREMIS is that there is no “opening” to link to the reference objects/files, e.g. colour interpretations/codes for pictures and similar for audio files.

Generally it is difficult to state whether the present preservation metadata is sufficient or not, mostly because of the relatively short period of time that has passed. Formats are not old enough to evaluate this. There might be a need for a thorough analysis of what metadata that is required for the future, i.e. quality assurance of metadata. Tools for metadata extraction can then be chosen based on this analysis.

There is no metadata search functionality in the DSM, and it is not clear if this is needed. Since metadata is copied to the bibliographic catalogues, these can be used for such searching. Normally, content indexes and content metadata, which is also extracted at ingest but not stored in the DSM but rather in separate index and search storage (see Figure 1), is used for searching.

4.4 Migration of content

The DSM repository is hardware independent, meaning that storage systems and media can be changed according to the constant development of better and more effective solutions. This has been carried out in practice, as NB is now (2008) on its fourth generation of storage technology since 1998. NB’s DSM system architecture is very modular and flexible. It is possible to change storage technology, suppliers, disk technology and such without changing the overall architecture.

Each change of storage system implies migration of all information from the old system to the new. This implies that *NB* needs *strategies for migration of content* when changing hardware storage systems. Although tools and procedures for such migration operations are known, the practical implications of applying them when it comes to huge amounts of data are not properly tested.

REPORT

It is a risk to loose data during migration because of procedure errors, technical errors, and, particularly, human errors because parts of procedures are manual. To minimize these risks one could do a thorough check on international initiatives and best practices and, perhaps, follow Trusted Repository Audit & Checklist.

NB uses a combination of standard software and OS commands for migration at present. Typically, UNIX commands like <mv file storage 1 storage 2> are used for manual file transfer from one storage system to another. Alternatively, the transfer function can be scripted as a mirror function in the OS; in this case the migration becomes automated and this provides certain protection against typing errors. The third alternative is to use a mirror function in the SAN (Storage Area Network).

Mirroring utilizes functionality is usually targeted at resilience (dual storage in the SAN) to copy objects from one disk to another within the same system. New disk technology can be introduced, including larger disks and file systems. However, mirroring to larger disks will leave the extra storage unused for a start, to be either filled by new objects ingested later or by more clever utilization of mirroring functionality, e.g. to mirror several old disks onto a new one.

The latter is an example illustrating that the migration process may need to be specifically tailored for technology changes. As another example, on a change of the file system or the file structure from maximum size 2 TB to 10 TB, five old “structures” should be copied onto one new. Such tailoring is assumed to increase the risk of failures, and documenting the migration becomes more cumbersome.

Both OS and SAN are assumed to apply protection mechanisms, and in particular mirroring is a process that must be “safe”. However, when the possibility of errors is larger than zero, even just slightly larger, the volume of migration jobs at NB indicates that errors may occur.

At present, an MD5 checksum is computed for each object. The checksum is verified when an object is read, not at the time of migration. If the checksum fails, the other two (tape-based) copies of the object can be tried. Verification at time of migration can be done by reading the object back from the new storage, then verifying the checksum to see that the new copy is not corrupted.

There is some uncertainty concerning whether one checksum algorithm alone gives the appropriate protection against loss of integrity. Adding another checksum using another algorithm may be an option but will imply processing overhead. Note that although MD5 is not considered cryptographically secure anymore, this is not a problem here. A cryptographic attack in the migration process is not a relevant threat; the risk is rather the following:

- Errors in the MD5 implementation or other software cause a corrupted object to pass undetected. A test bed to verify checksum functionality and its reliability to ensure integrity and verification is eventually considered to be a necessary task.
- There is a small, but perhaps significant, risk that a corrupted object yields the same MD5 checksum as the original object. Given that MD5 gives (pseudo-)random results, this risk can be computed as a function of the number of information objects, the likelihood of errors, and the likelihood of collisions in MD5.

The overall issue is that NB needs to prove (given some meaning of that term) that no object will ever be lost during a migration process. At the same time, the process must be efficient and as fast as possible.

The Danish National Library has conducted experiments and calculations indicating that data loss will occur in 114 years. A doubling of the data volume results in a decrease of 1/10 of the expected time to loss. Denmark has two copies in two different media types, while NB has three copies.

Overall, the main topics for improvement are probably the following, bearing in mind that NB's current situation is in fact the best practice based on the currently available knowledge:

- General guidelines for migration/storage should be developed, improving the current guidelines.
- Control mechanisms "outside the system" are needed.
- Should one have a general set of migration/conversion procedures or should one evaluate relevant software every time?
- Better time for planning migration processes and for completing the plan will contribute to better quality.

4.5 Conversion and other content management

At present, migration is the only content management operation performed; however further processing will be needed some time in the future. As examples:

- Sooner or later formats will become obsolete. For most objects, NB has a conversion strategy, meaning that AIPs must be converted from today's format to whatever is selected at that time in the future. This goes not only for the content but also for the metadata.
- It is possible that new technology may enable further metadata harvesting from AIPs, in addition to the metadata captured by the ingest process.

The alternative to conversion is an emulation strategy, meaning that the necessary software and hardware environment to process the format must be emulated on future equipment. For some material, platform specific computer games is one example, emulation is the only possible strategy as conversion must be deemed impossible. Emulation may also be chosen for heritage web dissemination, among others, because this allows showing documents written for specific browsers. For emulation, one has to trust the emulation software (and hardware) to work properly in order to give an authentic rendering.

It is essential to *develop ways to secure authenticity and validity in conversion processes*. One needs to look into the existing tools for verification and validation to test their quality and to establish where there may be gaps in the processes.

The difference between preservation format and access (read) format should be emphasised here. The latter has a shorter life time than the former. Access/read formats will typically be discarded as soon as they are not in use. For example, JPEG is used to read books nowadays. J2K compresses better though but browsers do not support it. As soon as a browser supports J2K or something else, JPEG will be disposed of. The discussions in this section are mainly relevant to the case where the preservation format changes.

Some examples of strategies may be:

- Ensure that the tools are 100 % trustworthy and that errors will not occur. Given today's state in hardware and in particular software, it is questionable if this is realistic.
- Where possible, and only for lossless (in some meaning of that word) conversion, automated tests may be used to check the result of the conversion. This may be checksums or tests similar to that applied (to samples) for the digitized books ingest process described above (convert from TIFF to J2K, then back to TIFF and check that the two TIFFs are equal). File size is one parameter. How about drawings and the like? Can appearance be compared?

- Apply two or more conversion tools in parallel and compare results (if possible) or store both (all) copies using number two as backup in case the first conversion tool has failed. Comparison may be on bit-level or by comparing appearance if that is possible.
- Loss conversion *should be* documented properly, but this is more at the stage of a hypothesis than an established practice. However, assuming that all conversions can be lossless is not reasonable, e.g. sound files where sampling rate, formats, stages of the conversion are stored in the header. Loss during conversion in other formats is not documented – the problem should be addressed (e.g. by registering parameters set up in a conversion programme).

Storing two copies instead of one has the unfortunate effect of potentially doubling the size of the object; assuming that all copies are stored in one AIP (meaning they share some metadata) and that all copies have approximately the same size. The same problem results if the original format (before conversion) shall be kept in the AIP along with the new format. At present, NB tends to believe (no firm decision taken) that the old format shall also be kept. At present, conversion results in an update of the *same object* (with the same ID) and does not generate a new object.

One reason for keeping the old format is that on a second (and so on) conversion of the object, one should convert from the oldest version that is intelligible, thus not propagating errors that may have emerged in intermediate conversions. Thus, one option is to preserve the original, the last version and the conversion statistics as metadata for all versions that have been in-between. This is done for sound files today where the original bonds are being kept. Ultimately, the decision on how many versions (and/or copies) to keep will surely be an economic question.

The conversion process must be documented in the metadata. The question that is potentially interesting is which objects have undergone a particular conversion process. NB may, perhaps, need the ability to go back and do the conversion again on the same objects but with new technology.

It is anticipated that large-scale conversion processes (possibly also other content management processes) must be integrated with the migration process. Running a separate conversion process covering the entire DSM (even if done on only one of the three copies) may be infeasible. The last migration process, which was low on the need for processing power (storage bandwidth was the bottleneck), needed several months to complete.

However, adding extra functions to a process based on mirroring (see above on migration) may be impossible or at least very difficult. If extra functions are performed, the objects must pass through a computer capable of (running the software necessary for) performing the conversions.

It is in general not possible to search for or otherwise pick only objects having a given format (say, all MPEG1levelII objects); however it may be possible to run a process covering a defined subset of the DSM (say, the radio broadcast archive) provided that the size of the subset enables the process to complete within reasonable time. Using the bibliographic catalogues to search up all objects of a given format may potentially be possible.

Similarly, it is in general not easy to find all objects that have been subject to a particular conversion process or that have been processed by a particular piece of software or hardware. However, it *is* possible to find objects with a certain property in the DSM.

4.6 What are the limits for migration and conversion?

Migration of a petabyte-size repository, i.e. one way or another copying the contents from one storage medium to another, takes several months using the technology available at NB today. The main limiting factor is the bandwidth of the storage media. The amount of information is increasing

exponentially, i.e. not only the volume increases but the rate of increase is accelerating. Given today's state, where migration is done every 3rd -5th year, one will sooner or later reach the situation where the previous migration is not finished when the next one is about to start. If more resource demanding processing, such as format conversion, is added to the migration process, the time for this situation to occur will be closer. On the other hand, advances in technology may increase the speed of the migration process and/or increase the intervals between the migrations.

To plan ahead for this situation, a formula should be devised to compute the time needed when different values are assigned to a set of parameters. This can also be used to identify the bottlenecks in the process in order to focus on these to speed up the process.

The relevant parameters are (at least): Size of repository, number of objects, number of copies, bandwidth for media access (read and write), computer processing speed, network capacity, lifetime of media, and lifetime of formats. Add to these secondary parameters as size of storage medium (such as disk or tape size), limitations on technology such as maximum size of file system, complexity of computing for migration, and complexity of processing for other operations (such as conversion).

This topic will be covered as part of a doctorate work within the LongRec project, rather than as a case study topic.

5 THE CHOSEN TOPIC

Develop general guidelines for migration/conversion providing "absolute certainty" of not loosing data, meeting the requirements of the trusted repository. Processes must preserve authenticity and validity in conversion and migration.

Solving this topic is a too tall order for the LongRec project. But the result of the case study should be a significant improved situation to NB.

Generally, most of the challenges at NB have to do with the enormous amounts of information NB deals with, both data volumes, vast numbers of information objects and large objects. For one thing, this means that everything must be automated except for sample checking. Another challenge is the heterogeneity of the objects – this is definitely not only text documents but also all kinds of (multi-) media material.

The guidelines should consider all or selected issues as identified in the previous chapter, and possibly more. Technical and procedural requirements should be developed, and if possible a pilot should be planned for testing the results.

6 BIBLIOGRAPHY

6.1 Relevant standards

MARC – Machine-Readable Cataloguing – is a format standard for the storage and exchange of bibliographic records and related information in machine-readable form. All MARC standards conform to ISO 2709:1996 Information and documentation -- Format for Information Exchange. See http://www.bl.uk/services/bibliographic/exchange.html

METS – Metadata Encoding and Transmission Standard – a standard for encoding descriptive, administrative, and structural metadata regarding objects within a digital library, expressed using XML Schema as specified by the W3C Consortium. The standard is maintained in the Network Development and MARC Standards Office of the Library of Congress, and is being developed as an initiative of the Digital Library Federation. See http://www.loc.gov/standards/mets/

PREMIS – Preservation Metadata: Implementation Strategies is an international working group that has produced a report "Data Dictionary for Preservation Metadata". The report defines and describes an implementable set of core preservation metadata with broad applicability to digital preservation repositories. See http://www.oclc.org/research/projects/pmwg/premis-final.pdf The PREMIS METS SCHEMA v1.1 is found at http://www.loc.gov/standards/premis/v1

OAIS – Reference Model for an Open Archival Information System – a technical recommendation on archive requirements to provide permanent or indefinite long-term, preservation of digital information. The recommendation establishes a common framework of terms and concepts. See http://public.ccsds.org/publications/archive/650x0b1.pdf or http://nost.gsfc.nasa.gov/isoas/

XML Schema (XSD) expresses shared vocabularies and allows machines to carry out rules made by people. Provides means for defining the structure, content and semantics of XML documents.

6.2 IT applications and software

DROID - Digital Record Object Identification – a software tool developed by the National Archives to perform automated batch identification of file formats. DROID uses internal and external signatures to identify and report the specific file format versions of digital files. It is a platform-independent Java application with a documented public API. See http://droid.sourceforge.net/wiki/index.php/Introduction

JHOVE – JSTOR/Harvard Version Validation Environment – a software tool providing functions to perform format-specific identification, validation and characterization of digital objects. It is an open source Java application. The standard representation information reported by JHOVE includes: file pathname of URI (Uniform Resource Identifier), last modification date, byte size, format, format version, MIME type, format profiles, and optionally, CRC32, MD5, and SHA-1 checksums. See http://hul.harvard.edu/jhove/

koLibRI – kopal Library for Retrieval and Ingest – is a library of open source Java tools developed by the kopal project for interaction with IBM's DIAS system. See http://kopal.langzeitarchivierung.de/index_koLibRI.php.en DIAS-Core interface specifications are available for Submission Information Package (SIP) and Dissemination Information Package (DIP) at http://kopal.langzeitarchivierung.de/downloads/kopal_DIAS_SIP_Interface_Specification.pdf and http://kopal.langzeitarchivierung.de/downloads/kopal_DIAS_DIP_Interface_Specification.pdf

NZ – Metadata Extraction Tool developed by the National Library of New Zealand – is an open-source software used to programmatically extract preservation MD from the headers of a range of file formats, including PDF, MS Word 2, MS Word 6, MS Word Perfect, Open Office, MS Works, MS Excel, MS PowerPoint, TIFF, JPEG, WAV, MP3, HTML, GIF and BMP. It uses the combination of Java and XML. See http://www.natlib.govt.nz/about-us/current-initiatives/metadata-extraction-tool/?searchterm=extraction

SIP Manager is an application (binary distribution) from Uppsala University for creating, transferring, and managing SIPs (Submission Information Packages) for archiving purposes. See http://wiki.epc.ub.uu.se/display/FV/SIP+Manager

6.3 Internal documents produced by NB

The following documents are in Norwegian and not publicly available:

- Documentation of DSM
- DSM core metadata (Kjernemetadata i DSM)

- QIC conversion documentation
- TIFF to j2k documentation
- Legal deposit radio documentation
- Deposit documentation on files from NB
- Technical metadata (Tekniske metadata i DigitALT) describes how NB uses PREMIS
- XML schema for broadcast legal deposit metadata

6.4 Some relevant projects and initiatives

Certification of Digital Archives Project http://www.crl.edu (note the Trustworthy Repositories Audit & Certification: Criteria and Checklist http://www.crl.edu/content.asp?l1=13&l2=58&l3=162&l4=91)

LongRec http://research.dnv.com/longrec

InterPARES http://www.interpares.org

PLANETS http://www.planets-project.eu/

DELOS http://delos.info/

DPE http://www.digitalpreservationeurope.eu/

DCC http://www.dcc.ac.uk/

PADI http://www.nla.gov.au/padi/

REPORT

APPENDIX: RESEARCH METHODOLOGY

Rationale for choosing the case study subject

The LongRec consortium is composed of participants with an express interest in the topics of the project. Most partners provide a monetary contribution in addition to the work hours they spend. Five partners have the role of case study partners (in InterPARES the term is "testbeds"):

- The National Library of Norway (http://www.nb.no) : Case studies in the READ (records transition survival) and FIND (long-term usage) areas;
- Brønnøysund Register Centre (public business registers in Norway – http://www.brreg.no) : Case study in the UNDERSTAND (preservation of semantic value) area;
- DNV Maritime (ship classification society – http://www.dnv.com) : Case study in the COMPLIANCE (preservation of evidential value) area, possibly also in the TRUST area;
- StatoilHydro (oil and gas company – http://www.statoilhydro.com) : Case study in either the TRUST or the COMPLIANCE area, or both;
- CSAM International (portal solutions for access to health care information primarily in hospitals – http://www.csam.no) : Case study in the TRUST area;
- The National Archive Services of Norway (http://www.arkivverket.no) joint with the Norwegian Ministry of Foreign Affairs (http://www.ud.dep.no) : Case study in the TRUST area;

As can be seen, all research areas are covered by case studies. Further cases may be added later on in the project. Case partners have been assigned to topics based on their own interest, with an additional criterion that the partner should be competent in the area and have a reasonably advanced solution in place (or under development). The rationale is that LongRec should focus not on solving today's problems for an immature case partner, but rather focus on bringing the long-term aspects in for an existing solution.

Research method

A case study is carried out by a research team together with a team from the case partner in question. The study is accomplished in several steps. After selection of the case partners one or two (typically) brief meetings with discussions with the key people of the case partner teams are conducted, preferably at the case partner's site. As a result of these meetings and additional e-mail/telephone communication, a short initial description (about three pages) of the case study subject is written by the case study partner.

After receiving the short case description, the research team prepares a number of interview questions tailored for each specific case but based on the InterPARES case study interview guidelines[12]. It is concluded that one standardised set of questions would simply not allow illuminating the problem areas and fully describing the situation for each initially outlined case study due to the differences in case study topics and the different nature of the case partner organizations. The relevant interview candidates shall be listed in the short initial case description.

[12] See http://www.interpares.org/ip2/ip2_case_studies.cfm

REPORT

The purpose of carrying out interviews is to *concretize* each case topic. Content analysis of the interview transcriptions is carried out after the interviews to identify the key areas of each case and to explore how interviewees' concepts might be linked to LongRec concepts. As an outcome of the content analysis, a list of all identified, possible research topics is derived. This list is fed into the general research activities conducted by LongRec as ideas for further research. Interviews are supplied by literature studies, typically material identified during the interviews (anything from internal documentation at the case partner, via standards and recommendations, to research papers), and usually also demonstrations of existing solutions. Email and telephone are used to clarify issues during the content analysis phase.

Then, the topics are discussed and evaluated jointly by the case partner and the research team. Usually several iterations are needed before concluding on one topic (may cover one or more of the research topics in the list) for the concrete case study. The topic is then specified by describing the present state, the desired state and the value its solution will give to the case partner.

At this stage, the first case study report[13] is produced, documenting primarily the list of topics and the single topic to focus further on. This report shall preferably be a public document. An important aspect of the report is to disseminate results to other LongRec partners.

Work on the case study then continues by a gap analysis between the present situation and the state of the art in research. This is described in a shorter, second case study report.

The next step in the work is to "solve the case" by detailing requirements, specifying necessary (work) processes, and identifying the technology needed. Changes to existing processes and technology must also be specified. Many, but probably not all, case studies will be concluded by trials/pilots testing both processes and technology. Results are documented in the third and final case study report, which is expected to be anything from 10-50 pages depending on the case study topic.

The timeline for the case studies varies from case to case and is not specified here.

o0o -

[13] This report is an example of such a report.