REPORT

THE NATIONAL LIBRARY OF NORWAY

LONGREC CASE STUDY: CHALLENGES IN SEARCH AND RETRIEVAL

DNV REPORT NO 2008-0543

DET NORSKE VERITAS

REPORT

Date of first issue:	Project No:	DET NORSKE VERITAS AS Research and Innovation
2007-10-14	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 .:	
Olga Cerrato, Lars Gaustad (NB), Jon Ølnes	The National Library of Norway/Nasjonalbiblioteket (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 search and retrieval performed by the library's relatively new (2007) general search solution based on FAST technology. The quality of search in huge data volumes contained primarily in the National Library's trusted digital repository is the scope of the case. Search is based on indexing the structured information like bibliographic catalogues and unstructured information like OCR output in scanning processes. In this study it is proposed that the search quality may be measured in such terms as completeness and relevance of the search results.

The main research question is whether the quality of hits returned by the search engine can be improved after conducting the analysis and subsequent tuning of the built-in ranking algorithms and client profiling.

The writings of this report are 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 and search, 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.

Report No:	Subject Group:	Date of this revision:	Revision No:	Number of pages:
2008-0543		2008-04-15	4	24

Report title:

The National Library of Norway LongRec Case Study: Challenges in Search and Retrieval

Report No: 2005-0543

REPORT

Table of Contents *Page*

Report No: 2005-0543

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 FIND (long-term usage) research challenge by investigating the challenges faced by the ***National Library of Norway (NB)***.

In 2007, NB launched a new, general search service[2] based on technology from FAST[3]. In a time perspective, improvements can be envisaged in two areas:

- How can search and retrieval in old text be improved?
- What can we do – if anything – to prevent or reduce problems for future retrieval in today's contemporary texts?

Search is based on indexing of:

- Structured information from bibliographic and other catalogues and from structural analysis of objects;
- Unstructured information from the entire content of textual documents (OCR in scanning process or analysis of objects that are submitted in digital form).

In the long-term perspective, challenges in search are related to alterations over time: language and terminology change, place names (toponyms) change, the meanings of words change and new words appear. How can search using today's terminology be mapped to terms that have gone out of use or changed meaning? What if the current term is "new" compared to the objects that the user needs?

[1] We refer to the project's web site http://research.dnv.com/longrec for a description of the research challenges.
[2] http://www.nb.no/sok/search.jsf
[3] http://www.fastsearch.com

At the same time, there are no resources to build ontologies/dictionaries that need manual maintenance (exceptions could be in particular, narrow areas of interest). Thus, the following topic is singled out for future work in the case study:

Test scalability volume/performance related to quality of hits by evaluating ranking algorithms.

Quality of search is measured in completeness and relevance. In a LongRec context, improving relevance and completeness over time is the important issue. Approaches can utilize search statistics: look at what terms the user enters and use this information to improve ranking for subsequent searches. The information gathered may consider age of the digital objects to deduce similarities between terms that create matches in "old" versus "new" objects.

Solutions can be piloted by tests on search in old material in DSM based on "new" keywords. If desired, a pilot can be narrowed down to a particular topic area, and in this context use of dictionaries and mappings can also be tested.

The idea of building ontologies by use of a search engine will be investigated in the future research in LongRec and this work may create input for the case study at NB.

The desired state for the long-term perspective is a system for indexing and searching that ensures reliable search functionality over time and assists the end user in the search process.

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

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 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 original medium, so the collection of NB covers everything from printed material, published music, broadcasts, film 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 on 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. By 2007, approximately one billion web pages have been downloaded.

NB shall receive and store all material broadcasted by the national and local broadcasters, although local broadcasting is collected solely on a sample level. NRK's[4] radio archives are being digitized and stored at NB. Digitizing of TV broadcasts is not yet prioritized for capacity reasons.

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).

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

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) – The Information Package, derived from one or more AIPs, received by the Consumer in response to a request to the OAIS (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).

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.).

Preservation Description Information – the information which is necessary for adequate preservation of the Content Information and which can be categorized as Provenance, Reference, Fixity, and Context information (OAIS def.).

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

4 SEARCH IN NB'S DIGITAL ARCHIVES – 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. The replaceable hardware system in use stores three copies of each object on two different technologies in two different localities. One copy is optimized for on-line access and is kept on (present technology) a RAID5 disk array. The other two instances are kept on tape robots. A SAM-FS file system is used.

REPORT

The architecture and implementation are according to the OAIS reference model[5] (ISO 14721:2003) shown in Figure 1 below.

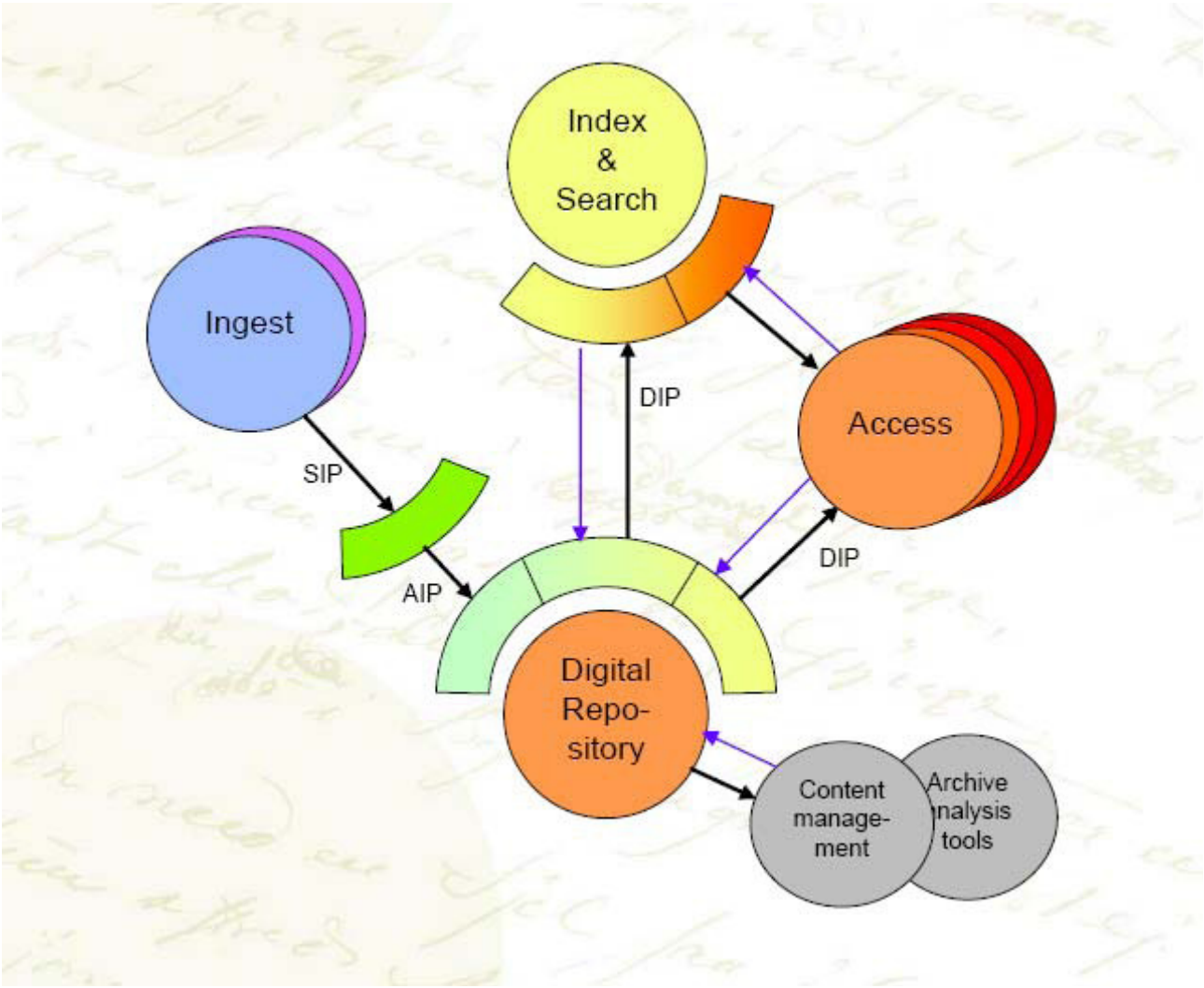

Figure 1: OAIS reference model.

As shown the model references three different packages:

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

The OAIS reference model defines six areas of concern/functional entities:

- Ingest
- Data Management
- Archival Storage

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

- Administration
- Preservation Planning
- Access

The information flow between the functional entities is shown in Figure 2 below.

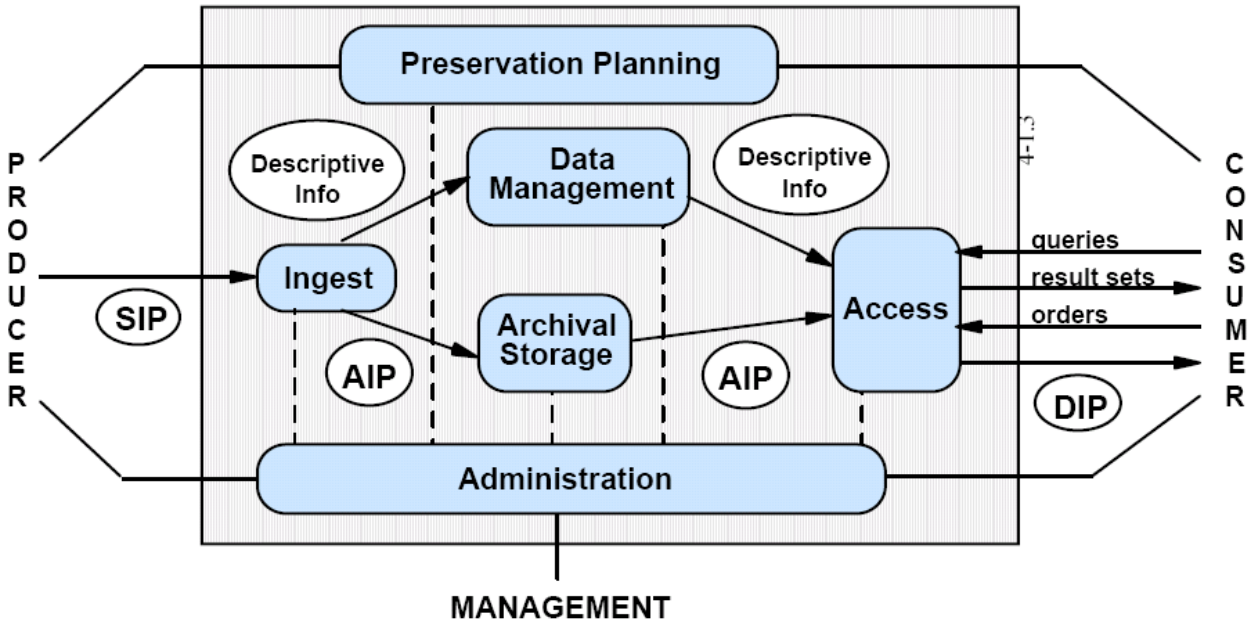

Figure 2: Functional entities in the OAIS model and their relationships.

This report studies the access part and more specifically search functionality in order to access content in the repository. However, the efficiency of search strongly depends on the job done in most of the other areas of concern and in particular at time of ingest.

4.2 Ingest

4.2.1 Ingest Processes

Building on the OAIS model AIPs are created with preservation metadata attached though use of extraction tools like JHOVE[6] and DROID. All processing to the file is registered as events in the PREMIS metadata schema.

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 JPEG2000 (J2K) (lossless compression).
- 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.

[6] See the bibliography chapter for a description of tools, standards and specifications.

- 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.
- Web-harvesting of the .no domain is run once a year.

The Legal Deposit Act does give the NB some control when it comes to legal deposit format and quality, but this is of course not relevant when it comes to the Web.

Not all information is stored inside DSM. While NB's strategy is to digitize all material, objects from several producers/owners are stored outside of DSM because the required work to place them inside has not yet been done. The objects are registered in the appropriate catalogues. The catalogue information may be searched but not the content.

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, indexing and metadata harvesting. Manual procedures are infeasible except for samples and possibly for some ingest processes with a fairly low volume. The different ingest processes and the different input data objects yield different levels of preservation and content metadata.

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/producer.

4.2.2 Digitizing Processes

Today, most objects are originally created in digital form (books, newspapers, photographs etc.) and one must assume that objects will increasingly be submitted to NB in their digital representation. However, submission to NB is still to a large extent on paper. Sound (radio, music), TV and film are also still to a large extent submitted in analogue forms. In addition, NB has large collections of old material where no digital representation has ever existed. NB's goal is to store all material in digital form in the DSM. Thus processes for digitizing *analogue* and *media specific digital* content will remain important for several years and it is important that digitized content is accessible to search engines.

Taking a book digitizing process as an example, the scanned book page makes two pictures (TIFF format) - one for each side. Then OCR (Optical Character Recognition, the software used is Docworks) captures text and the physical placement of the text on the page in order to deliver accurate hits during search afterwards. OCR quality depends on the quality of the originals. OCR reliability ranges from 100% to almost no recognition (e.g. rasterizing text on a colour background).

Automatic structure analysis (also done using Docworks) can only be effective when the OCR process yields a result of sufficient quality. The structure analysis supplies metadata for page number, chapter

[7] See http://www.daisy.org
[8] Synchronized Multimedia Integration Language specification, W3C Recommendation 07, 2001.

start, overall structure etc. together with text and its physical placement in the form of an XML export in the ALTO schema.

The software used is designed for some manual control but NB has no resources for this. Thus the result depends (almost) solely on the automated processes. Since future search in digitized material strongly depends on the analysis at this stage, there are clear challenges here.

Indexing (see 4.4) is done for the metadata (including structure analysis) and for the full text based on OCR.

There are many sources of metadata (see 4.3) in the scanning process:

- The first part of the metadata capture is done prior to digitizing; all physical objects are registered in (bibliographic or other) catalogues and metadata is fetched from the source catalogue of the object to be digitized. In addition to internal catalogues, NB also refers to completely external catalogues like BIBSYS[9].
- Then, the scanner adds information on the type of the scanner, resolution, etc.
- In the scanning process, metadata on scanning is generated for each scanned page.
- The automated structure analysis adds content metadata.
- Preservation metadata is added by use of tools such as JHOVE and DROID.

Finally, format conversion of pictures from TIFF to JPEG2000 (J2K) takes place, catalogues are updated, and the object is stored in DSM.

Digitizing is used for many types of objects, not only books. Other printed objects have more severe challenges with respect to structure analysis, e.g. posters where analysis of the text is difficult. Paper photographs are particularly challenging as discussed in 4.3.3.

Digitized broadcasting material (radio, eventually TV) should as far as possible have metadata in time of creation/publication, such as source, titles and other information on programmes (can be linked to structural metadata). Content metadata is not included today but one could envisage speech to word processing as a mean to enable search in content.

At present, OCR and structural analysis are only done at ingest. It may in principle be possible to do a new analysis of objects that are already in DSM, if e.g. improved tools become available.

4.2.3 Digital Submission

One may argue that the problems of digitizing analogue material are temporary only. Eventually, all old material will have been digitized and all new material will be “born digital”. However, ‘temporary’ in the context of NB probably means ten years, perhaps, twenty years or more. Changing the processes for all types of content and all actors to submission of digital versions of the content (where relevant in parallel with submission of the “hard” version of e.g. a book or a DVD/CD) is bound to take time.

For publications such as books, newspapers and magazines digital submission should enable far better searching since content indexing will be 100 % correct in comparison with OCR.

In principle, structural analysis should also be a lot easier than for digitized material; however unless there is useful structural metadata submitted, one needs structural analysis similar to the OCR case to

[9] http://www.bibsys.no/wps/wcm/connect/BIBSYS+Eng

single out title, headings, body text etc and this may depend on potentially complex analysis of tags, styles or other formatting information.

One reason is that today's usual situation is that only small amounts of metadata follow digitally born objects. Information about what equipment was used for production as well as characteristics and quality of equipment is frequently missing. This might be necessary for preservation but this may be a temporary problem (in some meaning of the word temporary).

Naming of digital objects may be more or less arbitrary, e.g. photographs may be named randomly by photographers/authors.

Note that the amount of metadata submitted with the content may vary:

- As mentioned above, two major newspapers submit PDF-files today with no metadata but with filenames that relate to date, version, zone, section and page numbering only. In such cases, structural analysis is similar to analysis of digitized material.
- Legal deposit radio is an example of digital submission with attached metadata in the form of an XML export of NRK's production database covering one day of descriptive metadata.

For other objects, new sources and types of metadata will emerge. Consider for example a future digital camera with GPS and possibly gyro/compass to indicate direction; this may allow connecting photographs to geographical co-ordinates and even place names. Time, identification of camera and other information are already included in the metadata of most digital photographs.

Increasingly, material will be published only in digital form and not on any tangible medium. Some such publications may be printed or otherwise "materialized" by a consumer but in other cases digital publishing is fundamentally different from physical publishing. A printed book has some linear ordering of the content, while a web site consists of pages and links that are not "linear" in the same way. A physical medium defines the borders of the content while these borders are hardly defined for a digital publication with links, not to mention a web-page dynamically built from content from different sources (e.g. advertisements separate from other content).

Today's web submission method is not submission as such but active web harvesting of the .no domain once a year. Web-harvesting poses some particular problems:

- Speed and scaling of the current process is questionable.
- 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.
- Intellectual property rights and privacy issues block general, on-line access to the archived web content.
- There is insufficient control over formats and very limited control of metadata harvesting.
- Sites that are subject to access control are not covered.
- Web pages that are dynamically created e.g. based on query formulation are not covered.

There is a need to define arrangements for active submission of content, e.g. authors publishing on Internet. Mechanisms such as RSS feed may be used to detect updates on selected sites. Scheduled download processes may be run more frequently for selected sites, e.g. daily or weekly. Targeted download processes may also be run in conjunction with particular events, e.g. elections, harvesting relevant sites (static or dynamic identification or based on search) quite frequently for a limited period of time. These topics are not discussed further in this report.

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. An idea is to use metadata for referral between objects in DSM to organize this kind of material.

The latter identifies the issue of what constitutes an object in the repository. To what extent should it be possible to follow links between objects, either the original links or replaced by relationships between objects in the repository?

4.3 Metadata

4.3.1 Preservation Metadata in DSM

Metadata may be divided into preservation metadata (necessary or desired in order to preserve an object over time) and content metadata (necessary in order to characterize the information in the object, such as structural analysis and indexing). Both types of metadata may be used in search.

At ingest AIPs are created with preservation metadata derived and attached though use of extraction tools like JHOVE[10], DROID and NZ and static XML-files to populate a METS conforming schema, built on the APSR/NLA[11] METS schema. The APSR/NLA schema implements PREMIS schemas and other defined METS schemas like MODS, MIX, LoCs AudioMetaData.xsd and VIDEOMetaData.xsd. 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.

All processing to the file is registered as events in the PREMIS schema. Where relevant, preservation metadata is also exported to NB's catalogues (see below). The AIPs stored in DSM consist of the information object and an XML file containing the METS metadata 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.

NB has specified its *Core Metadata* that in the context of 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 metadata – called *relations* – that indicate how *files* belong together. E.g. NB digitalizes 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.

Authorizations and access restrictions can be represented in METS; however this is not in use at NB today. Some material is 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.

[10] Refer to the bibliography chapter for a description of tools, standards and specifications.
[11] METS profile developed by the National Library of Australia (NLA) through the Australian Partnership for Sustainable Repositories (APSR).

Note that a current problem with PREMIS is that there is no "opening" to link to reference objects/files, e.g. colour interpretations/codes for pictures that may be necessary to understand the files, and similar for audio files. The PREMIS development team is considering implementing this for the 2.1 release of PREMIS.

4.3.2 Catalogues and Preservation Metadata

All physical objects at NB should be registered in (bibliographic or other) catalogues containing identification of the object and associated metadata. NB also refers to completely external catalogues like BIBSYS. The MARC format is predominant for metadata in catalogues covering books and periodicals although different formats are in use. For material born digital, catalogues do not necessarily exist (e.g. not for web harvesting). This depends on the submission process for the digital or physical material.

Cataloguing is done according to Norwegian cataloguing rules (based on the Anglo-American Cataloguing Rules II) for most documents, for some document types according to other local, national and/or international rules, especially for audio and sound recordings. Classification numbers (Dewey's decimal classification, Norwegian version) are assigned to a majority of the records; keywords are used for some collections.

MAVIS is the catalogue used for audiovisual material as well as some other media types. MAVIS includes both digitized and born digital objects.

Cataloguing in general is not meant for electronic procedures, only manual procedures. Traditionally, a lot of resources were used on book cataloguing while other media were neglected. Sound and pictures have a shorter registration history and registration practices with varying quality. There are initiatives (e.g. by Library of Congress in USA) at minimizing efforts in book cataloguing while improving cataloguing for other material than books (sound, pictures etc.).

If information about an object in DSM is contained in a catalogue, the catalogue information shall be updated by a pointer to the DSM object. Additionally, all metadata from the ingest process should be copied to the catalogue, provided that the catalogue can handle the information.

When an object is digitized, the MARC information from the catalogue is also included in the METS schema that is associated with the object in DSM. Conversion from MARC to METS is lossless.

4.3.3 Catalogues and Search

The following metadata should be present in the (bibliographic or other) catalogues outside of DSM:

- Metadata that describes the content and that can be used for indexing and search;
- Metadata describing (intellectual property and other) rights to the material;
- Metadata describing availability and access rights to the material.

The availability of this information depends on several sources:

- Structured metadata supplied with the object at submission and/or entered into NB's internal catalogues;
- Structured metadata available in external catalogues such as BIBSYS;
- Unstructured metadata that can be derived from the object during the ingest process, notably from OCR and automated structure analysis.

There are challenges in search related to improvements in metadata at NB and external parties. The problem is that metadata sources are heterogeneous and when these are brought together the results are inconsistent; information registered in one source is not present in the other. The reasons could be many: a source deemed it unimportant to have a type of information or a pure omission was made, or the automatic structure analysis of objects of different types yields different results. The technical integration towards catalogues in different systems and with different interfaces also causes some complexity.

Example: theme, topic or subject descriptions that are important for search (text, picture, radio, etc.) are not always registered; even if they are registered, this can be done in various ways. If information on a subject is missing, then the subject field is empty or the field does not exist at all.

Paper photographs are perhaps the most challenging objects. It is difficult to extract metadata automatically from photo archives (without any related context). There is little additional information (catalogues) that could provide any basis for metadata. Postcards may have some text that can be used as the metadata basis, but this is not always the case.

4.3.4 Content Metadata Representation

While *METS/PREMIS* is used for the preservation metadata (see 4.3.1), ***MODS*** is chosen to connect metadata from different source catalogues into a search service. MODS and METS are used by most other national libraries and are fronted by the Library of Congress. MODS is an XML schema that covers the old MARC format and enables the creation of additional description parameters. MODS and METS are linked in that MODS can be used as an extension schema to METS. MODS has descriptive metadata some of which are also present in METS.

MODS can point to external metadata, but this option is not in use. Index changes are done rarely, and if performed the trigger would be changes in NB's needs rather than changes in the standard.

Dublin Core is an alternative to MODS and was used in the previous version of search functionality; see for example "Kulturminne Ekofisk"[12]. The experience has shown that Dublin Core is too simplistic for search, i.e. has too few fields. *MARC* could be another alternative as this is a standardized format for bibliographic catalogues. However, MODS is better suited than MARC for representing other types of objects (audio, photographs etc.).

Thus the decision is to convert all source catalogue formats to MODS. The conversion from MARC to MODS is lossless. *MARC Exchange* is XML representation of the old MARC format. There is a Norwegian version of the MARC format, *NORMARC*. At the time of writing (March 2008) there is a planned transition from NORMARC to MARC21 going on. MARC has, among other things, an option of entering partial time information – e.g. approximate year, or date without a year. This is preserved in MODS.

Content metadata (e.g. actors, producers, etc.) and preservation metadata for audiovisual material are registered in ***MAVIS***, which is also indexed for searching.

Metadata is in Norwegian. Foreign language transcriptions are done in a number of ways. In MARC catalogues, catalogue posts use the book language normally, but metadata is in Norwegian (transcribed from the foreign language for the title with tags in Norwegian).

[12] http://www.kulturminne-ekofisk.no is a web-based exhibition of cultural heritage from the Ekofisk oil field in the Norwegian part of the North Sea.

There is a separate bibliography for material in Sami and all Sami-related material. This area is prioritized at NB and includes web-harvesting as well (more than 13000 documents).

4.3.5 Metadata Quality Issues

All concerned, NB has in general quite good control over the content metadata. Some issues are noted:

- As stated, there are differences in metadata obtained from different sources. Some sources contain free text, i.e. they are not completely structured. NB is trying to reduce the number of metadata sources and improve the quality of the sources.
- More metadata is needed for digitized material than for the paper version and this may cause gaps. This applies to sound, pictures etc. as well.
- Historically, some material has very little metadata – systems that were random, without systematised metadata. This is the case of some posters and photographs.
- Photographs are examples of digital objects that may not have meaningful names assigned; they are frequently named randomly by photographers/authors.
- Metadata for pictures could be much improved if information on the content of the picture was available, such as place name, the name of the plant, animal, etc.
- The automated structural analysis introduces some errors. An example of a mistake is a pricing list interpreted as a table of contents. Only a small fraction of the input can be manually checked.

NB could clarify metadata quality requirements further. This could aid direction of resources towards the most important issues for better quality, including more precise requirements for submitted metadata and metadata in external and internal catalogues.

Automated structural analysis might be improved by carefully applying manual checks but then the cases to check must be singled out by the automated processes. If manual checks are applied, the resources must be carefully directed at cases where there is a clear gain.

The metadata do not capture all information. For example, what calendar is it meant in Jewish date versus Norwegian date? This is an example of information (calendar) that is not a part of the metadata, but that can influence the search logic.

The severity of the quality issues is determined by the future consequences of metadata shortage. An analysis of these consequences would then guide the amount of resources that it is sensible to spend on quality improvement. Consequences are not necessarily severe but the main issue will typically be inability to identify an object during search.

4.4 Indexing

Indexing (the last stage in the production phase) is based on FAST's[13] solution and covers content metadata from different catalogues as described above as well as full text of textual objects as far as possible (e.g. limited by OCR quality). The documents are indexed 'as is', structured metadata from catalogues and full-text from OCR or born digital objects. All structured metadata are converted to MODS and fed into NB's public search service.

No semantic analysis or pre-processing is done before indexing. Each document is fed to the index as an individual XML record. Each record contains a MODS part for metadata and a METS section describing the physical and logical structure of the document (including the full-text content). The

[13] http://www.fast.no/

METS section is so far only implemented for digitized books. The MODS metadata are mapped to individual index fields so that it is possible to search specifically for a title, an author, a subject term, a date etc. In addition, the complete MODS record including the document's full-text (with structure) is stored in a so called *XML scope* field in the index. This enables querying the full-text content of the document. When querying, the result returned carries information about where in the document the match was (e.g. what pages, which paragraphs in a book). In addition the scope search functionality allows querying a record's individual MODS element values and attribute values directly (not utilized today).

Descriptive types of data like names registered/spelled in different ways and subjects/themes registered in different ways are a challenge. The same concerns book titles that are spelled according to older language styles. Shortenings and abbreviations are also a challenge. Occasional errors in OCR introduce mistakes and further undesired variants.

There is no reference to language changes. As an example, the Norwegian words "åndssvak" and "evneveik" were in common use for many years to characterize "mentally handicapped". Both words are now considered derogatory and replaced by the term "psykisk utviklingshemmet". A search on one of these terms should preferably display hits that contain all of them.

For non-textual objects, the structured metadata is usually all that is available. Photographs are perhaps the most challenging digital objects illustrating lack of metadata. This situation should improve with increased submission of digital version of photographs with metadata from camera and possibly other sources.

Should it be possible to search in other ways than text-based, e.g. find all pictures resembling an input picture? This is hardly relevant for today's NB users but might be so in the future.

4.5 Search and Access

A simple search query entered through NB's public search service[14] queries several index fields like name, subject and title as well as the full-text part of the *XML scope* field. The ranking of the hits returned depends on in which field(s) the query term was found. For instance, a match in the title will give a higher rank than a match in the full-text content. Some parameters can be adjusted on line but most of the rank settings must be set prior to building the index. Several 'rank profiles' may be defined prior to building the index, for example per user group but the users cannot set them themselves. The rank profile to use may be chosen at query time. The present solution offers navigation and limiting functions based on digital/non-digitised, type of media, date, places, end even Dewey.

The search interface offers unrestricted search in openly available material only. There is also authentication and access control solutions that enable search even in some restricted material.

Note that even in the presence of the new, general search service, users can still conduct search in individual catalogues. Search functions in separate catalogues can be better because the interface is then adjusted to the specificity of the catalogues and this can give better search hits than the general solution. Some context info can be specific and present only for certain catalogues and be missing in the general solution (but this is then deemed as non-essential).

[14] http://www.nb.no/sok/search.jsf Note that the Internet archive (.no harvesting) is at present not open to general search. This is due to unresolved issues in intellectual property rights and privacy.

4.6 Other relevant search aspects

NB's present solution has the option to use search statistics to improve ranking, e.g. look at what terms the user enters and use this information to improve ranking for subsequent searches. This is not implemented and there is no tuning in the present solution.

Currently no pre-processing of the queries is done: the query term entered is the exact term searched for. Search is based on the combination of search key words entry and navigation, where navigation is dynamic, based on search results. The hits are generated only for 100% similarities, not for alternative transcriptions. Glossaries could be considered in such cases.

No taxonomies are being used in the present solution. FAST's software extracts names, site names etc. and FAST has lists (aliases) in order to index that a term is a name. There is a possibility of adapting these lists so that they would refer to external sources as well (this option is not in use today).

Scalability=volume/efficiency. FAST's search engines scale well with respect to both data volume and search speed. If necessary, several computers can be attached.

There has been no substantial analysis done on the search results quality. A clear advantage of the present solution is that it ranks hits according to proximity of search terms in the material. If search terms occur in the same paragraph of a document, this yields a higher ranking than just having the terms present somewhere in the entire document. Searches can be performed across all types of content and any type and number of sources. The quality of search can be measured in completeness (everything is shown in the results list) and relevance. Search is definitely not complete and contains some "noise"; the exact quality has not been analysed.

There is no metadata search functionality in DSM itself, and it is not clear if this is needed. As described earlier, metadata is held in the relevant catalogues, and these are used for such searching.

In the long-term perspective, challenges in search are related to changes over time: language and terminology change, place names (toponyms) change, the meanings of words change and new words appear. The problem is relevant both when we look back and when we look into the future. NB can already experience problems with search and retrieval in older texts (from old digitized books and newspapers). Problems are foreseen when searching in contemporary text in the future. In essence:

- How can search and retrieval in old text be improved?
- What can we do – if anything – to prevent or reduce problems for future search and retrieval in today's contemporary texts?

Can semantic analysis be used when indexing to ensure better quality in search and retrieval? Is it possible to search in different, additional "semantic registers" in order to have better precision in search and retrieval of documents?

As an example, different versions of place names and different names for the same place could be listed. The Norwegian Mapping Authority (Norwegian: "Statens Kartverk") has a register of Norwegian place names that could be used as an external reference to different naming forms. Access to the register requires an agreement with the mapping authority and one needs to ensure that historical information is kept in the register. A test could be how to obtain a hit on a photograph depicting some town where different naming forms are used: one in a search term and another in the metadata of the photograph (e.g. "Trondheim" is searched for while the photograph is tagged with "Nidaros"). The photograph's age information could be used in order to search for a name that was in use at that time.

Mapping between the two main Norwegian language forms (Norwegian "bokmål" and New Norwegian) or even including dialects is a related topic. But in a long-term perspective the "versioning" of terms over time is probably more interesting.

Possibilities for search improvements are:

- Look at terms that appear often in proximity or take into account the user's search history.
- Use hits in documents in order to find other relevant terms – this is a new search in a limited set, but not the same as a word list.
- Foreign language transcriptions are done in a number of ways. The hits are generated only for 100% similarities, not for alternative transcriptions. Glossaries can be used in such cases.

Ideas on the algorithm level have a more generic relevancy than software tuning. As an example, there have been discussions in the LongRec project on use of search technology to build semantic understanding. The search engine could create correspondence between terms without use of pre-defined word lists.

5 THE CHOSEN TOPIC

Test quality of search results, related to relevance and completeness, in huge volumes by evaluating ranking algorithms and client profiling.

As stated above, no substantial analysis has been done on the search quality. Quality of search may be measured in completeness and relevance; these elements should be analysed.

In a LongRec context, improving relevance and completeness over time is the important issue. Approaches can utilize search statistics: look at what terms the user enters and use this information to improve ranking for subsequent searches. This functionality is at least partly in place but is not used in the present solution. The information gathered may consider the age of objects to deduce similarities between terms that create matches in "old" versus "new" objects.

Solutions can be piloted by tests on search in old material in DSM based on "new" keywords. If desired, a pilot can be narrowed down to a particular topic area, and in this context use of dictionaries and mappings can also be tested.

The idea of building ontologies by use of a search engine will be investigated in the future research in LongRec and this work may create input for the case study at NB.

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/

MODS – Metadata Object Description Schema is a schema for a bibliographic element set that may be used for a variety of purposes, and particularly for library applications, http://www.loc.gov/standards/mods/.

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. Version 2.0 is due in the spring of 2008.

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/

OAI-PMH - Open Archives Initiative Protocol for Metadata Harvesting v 2.0 facilitates cross-platform searches on internet resources, http://www.openarchives.org/OAI/openarchivesprotocol.html.

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

Docworks from CCS (Content Conversion Specialists) is a technology that helps through automation to disassemble electronic, paper, microfilm, or microfiche documents to its constituent parts and create searchable content while tagging structural and semantic metadata, http://www.ccs-gmbh.de/en/digitization.htm.

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/

FAST ESP, http://fast.no/thesolution.aspx?m=376

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/

REPORT

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 the COMPLIANCE area;
- 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[15]. 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.

[15] 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[16] 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 subsequent 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 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 -

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