REPRESENTATION OF DIGITAL MATERIAL PRESERVED IN A LIBRARY CONTEXT

Eld Zierau
The Royal Library of Denmark
Dep. of Digital Preservation
P.O.BOX 2149
1016 Copenhagen K, Denmark

ABSTRACT

This article explores preservation of digital material in a library context with a focus on logical object modelling that takes both preservation and dissemination into account. The article describes normalisation of data expressed via a logical object model. This logical object model is designed to support the requirements for joint preservation and dissemination. Additionally the article includes a suggestion for a possible implementation that respects the logical object model.

Formulation, of the requirements and possible implementation for a logical object model, is based on observation of current trends, as well as results from a research project on preservation strategies for libraries. The research project has been carried out at the Royal Library of Denmark, and it is based on a case study of a 10 year old web application containing the Archive of Danish Literature. The formulated requirements include e.g. requirements for many-to-many migration in preservation and requirements for homogenous navigation and social networking in dissemination.

Many of the described observations and results have parallels to other types of material. These parallels are partly described, and thus the results can be used as a contribution to development of systems and strategies for preservation and dissemination in the new decade and beyond.

1 INTRODUCTION

This article explores digital preservation in a university and national library context where preservation must go hand in hand with dissemination. It focuses on the object modelling aspects to represent a normalisation form that supports future functional preservation as well as dissemination. Functional (logical) preservation here means preservation of a digital object to ensure that it remain understandable and usable on a long term basis. The study is a result of a research project at the Royal Library of Denmark (KB), the goal of which is to investigate preservation strategies in a library context.

The hypothesis investigated is that it is possible to reuse and normalise existing data from digitisations (10 years or older). If this is the case, it will be economically beneficial to preserve the normalised data in the sense of preserving the investment of the earlier digitisations. The results of exploring the hypothesis will influence the future normalisation of data as well as preservation and dissemination strategies.

The research is based on a case study of the Archive of Danish Literature (ADL) system. ADL is a web-based framework constructed at the start of the century. ADL is mostly limited to books, book collections and book metadata, but parallels to other types of material can be drawn. A separate part of the research project investigated whether the original digitised ADL was worthy of preservation for future use (study part 1) [5], which the study found to be the case. The other part of the study is the one presented here. This part will only look at the normalisation and logical object modelling aspects for the digital material and their data structures.

In our view preservation and dissemination are highly interrelated. This leads us to assume that they must be managed jointly on a day-to-day basis regarding ingest, access and maintenance, as illustrated in the Figure 1. The terms used here are defined in the OAIS reference model[1], unless an explicit definition is given.

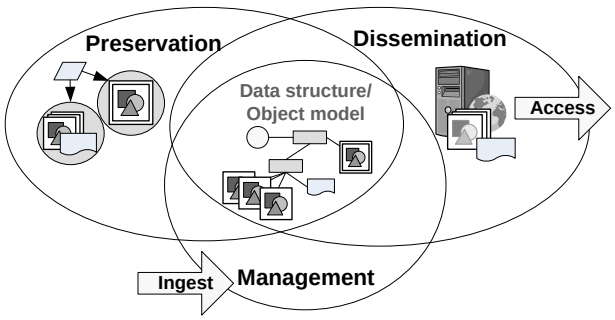

Figure 1. Preservation and dissemination interrelations.

The background for this view is that libraries have an obligation both to preserve and disseminate material. This fact challenges the demands on preservation, where material in many cases must retain a short and efficient route to dissemination through fast access by the public or researchers and in a user friendly way. Both dissemination and preservation demands are under constant challenge as a result of technological evolution. New requirements emerge such as representations to new media e.g. mobile devices, representations in new form e.g. e-books[2] or high resolution images, and representation information via social network communities[3] [1]. This means that digital material becomes more inhomogeneous with new representations. Furthermore, the need for different

[1] OAIS (Open Archival Information System). 2002. ISO 14721:2003.
[2] See http://en.wikipedia.org/wiki/E-book
[3] As define on http://en.wikipedia.org/wiki/Social_network_service

preservation levels becomes more apparent. Ten years ago, the focus was primarily on digitised books, while we today face challenges with e.g. contents from a PC of a deceased author, internet harvests, emails, and digitised images from deteriorating negatives [2].

The purposes and goals for dissemination and preservation are different. Their interrelation means that the requirements for dissemination need to be taken into account when we formulate the long term preservation strategies. Furthermore there are requirements to allow for data migration into preservation formats with different storage characteristics. Migration will here mean modification of the digital objects to ensure permanent access to these objects. The storage characteristics can be: how much storage space the format requires, or how different parts of a logical object e.g. a page image, are stored with different confidentiality levels and different bit preservation levels [7], i.e. different bit safety levels ensuring that the actual bits remain intact and accessible at all times. Most of these requirements must be taken into account when we define an object model for normalised data.

Before we can describe an object model for normalised data, we will list the relevant dissemination and preservation requirements based on the case study, the experiences gained, and the relevant results from study part 1. Some of the requirements will relate to an actual system implementation. This article will therefore include a description of a possible solution for digital object management systems (DOMS) that can support workflows of ingest, ensuring preservation and dissemination of the digital material of a library. The possible solution description is based on results from a DOMS pre-study at KB carried out by joint forces from the Digital Preservation Department and the Digital Infrastructure and Services department at KB.

2 CASE STUDY: THE ADL SYSTEM

The ADL System is used as a case study, in order to study new requirements for dissemination and preservation that emerged as a consequence of the technical evolution in the last decade. The case study is interesting because it reflects a system built on the basis of technologies from the start of this century. The case study gives us indications of the challenges to take into account when we consider a future DOMS, regarding present requirements, and regarding trends that should be addressed for future requirements. Although the ADL system is a case study covering specific materials, the indications will have parallels to other types of material. When the requirements are specified in the next section, such generalisation will be made where possible.

2.1 Short Description of ADL

The ADL system was developed by KB together with "Det Danske Sprog- og Litteraturselskab" (DSL) which publishes and documents Danish language and literature. KB developed the framework, while DSL selected literary works to be included. The system is a web based dissemination platform for digitised material from the Archive for Danish Literature. Today it contains literature from 78 authors represented represented by over 10,000 works of literature (defined as a work by an author that can represent itself without other context, examples are novels, poems, plays). ADL additionally contains author portraits as well as 33 pieces of music (sheet music) and 118 manuscripts. The publication framework is still available on http://www.adl.dk/.

The structure and design of the underlying ADL database is based on book pages, authors, their literary works and the period when the authors were active.

Since ADL was designed a decade ago, its navigation and search facilities along with design of data structures are old-fashioned compared to the possibilities of present technology. Although ADL has served as a good application, it now needs renewal which will partly be specified on basis of the research results.

2.2 Experiences from ADL

The ADL system does presently offer separate views of book pages in three ways based on three different digital representations of the pages, but there are no relations between the views. The views are: a 4-bit GIF image, a pure text representation, or a page can be downloaded as a PDF file containing the page image for print.

The data structure is highly dependent on pages, which gives several challenges. The structure of page images in a book is specified in a TEI-P4[1] LITE XML. The XML is uploaded to a database which is used for dynamic generation of HTML pages. The page number is used in the name of the related page files with page image and encoded text. This eases application coding of references to different representations of a page in GIF, text or PDF, but introduces a number of challenges. Firstly it challenges maintenance if page numbering needs to be corrected, not only should the file name be changed, but all references from e.g. citations via hardcoded URLs will need update as well. Another related challenge is that there can exist different versions of a page image. For example, ADL had a copyright restriction on illustrations appearing as part of a page. This restriction was only enforced within a certain period, thus two versions exists for such pages, both in the GIF image, and in the PDF derived from the original TIFF image. File names with page numbers will also cause problems for functional preservation that are similar to the problems of preserving web archives[2].

The navigation and search facilities depend on older technologies and the data structures. Limits became apparent in particular for navigation, when sheet music in PDFs with JPEG images were added (originally digitised for another purpose), and when manuscripts represented in JPEG files (for better dissemination of colours) were added. One problem was that this new

[1] TEI (Text Encoding Initiative).

[2] See e.g. "Long-term Preservation of Web Archives – Experimenting with Emulation mad Migration Methodologies" on http://netpreserve.org/publications/ NLA_2009_IIPC_Report.pdf

material is not viewed as literary works and therefore did not fit in the original navigation structure. Furthermore navigation of sheet music between pages is different, since they are fully represented in a PDF file.

Inclusion of additional material information in ADL has made the lack of referencing possibilities apparent. Examples are reference to other resources on external web-sites, or Danish translations for books written in Latin. The original ADL data model was not designed for these inclusions and they are therefore not logically integrated in ADL, e.g. the translations are hard to find and the relation to the book is not obvious.

A rare challenge in ADL occurred when a literary work, in the form of a novel, was added. The challenge was that the novel was represented in two volumes. The solution was to represent the two volumes as one book in ADL, with one XML file for both volumes.

ADL has an option for users to send an error report on errors in the OCR text. A challenge here has been to have dedicated time to handle the error reports, which are handled manually. Furthermore, the current ADL system does not have automatic version control on changed text, thus the changes can be hard to track.

Presently, the ADL is only preserved as a part of the Danish web archive. That means that the only data preserved is the data visible on the internet, which does not include e.g. special encoding of texts. Further actions for preservation await the research results.

2.3 Relevant Results from Experiments

In connection with the study part 1, we have done experiments involving two re-digitisations. Some of the results from these experiments also influence the normalisation considerations, therefore we here provide a short summary.

The re-digitisation was carried out in two places and with two different approaches. One carried out a mass digitisation including new scanning of the books (referred as SC1). Another used an approach similar to the original ADL digitisation (referred as SC2).

A conclusion from study part 1 was that the original ADL scans were worthy of preservation. There were, however, cases of missing pages in the ADL scans. The missing pages were mostly blank pages or pages with editorial information, but in one particular case, the missing pages contained parts of a poem. This gives an example of a case where we would like to add page images from the new SC1 scan to the existing ADL.

Another conclusion was that the original ADL XML encoding was worthy of preservation, but additional results from SC1 and SC2 should be added and preserved as well. The additional results were the encodings for the missing pages, and the marginal notes which originally were left out in the ADL encodings.

Updating the encodings challenges the representation. One challenge is that the encoding results differed due to the different encoding formats. The differences are both in coverage and in type of XML tree structure. The ADL and SC2 XML are given in TEI-P4 per book and the SC1 XML is given per page in ALTO[1]. Positions in ALTO from SC1 refer to SC1 scans, while it is the ADL scans that are preserved. Thus if positions are added for future referencing mechanisms or creation of searchable PDF, we will need to produce this information based on the ADL scans. Lastly, the encoding of marginal notes is interesting, because the SC1 XML marking notes via positions was the most precise result. In the SC2 XML notes were marked notes with reference to a full paragraph, which is not precise.

3 REQUIREMENTS

On the basis of our knowledge of growing demands, experiences and experimental results, we can now describe the requirements for dissemination and preservation. These requirements can be applied for book collections in general, and for other materials.

3.1 Requirements for Dissemination

The technological evolution of the last decade has opened many new dissemination possibilities. For example, faster internet connections have made it possible and more common to have videos and high resolution images as part of web material. Digital born material like e-books is becoming more common. More advanced presentation in websites is appearing, e.g. synchronised representations with annotations possibilities[2]. Consequently, the requirements for the ADL application are increasing in accord with these new possibilities. The information we want to disseminate has evolved as illustrated in Table 1.

Present ADL dissemination	Extra desired dissemination
Book page images (GIF-images, text, PDF download)	Other book manifestation of book item
Author citation	Content segments
Author description (picture, period, important dates), Period description	Thematic ontology Timeline with literary works
Sheet music & manuscripts	Other related material
Overviews (list of literary works, author list, period)	Time line, thematic ontology, student material, etc.
Error reporting option	Social network community (OCR correction, annotation, quiz etc.)

Table 1. Present and future dissemination.

The contents of Table 1 is based on generalisation of the current contents, on current technologies as

[1] ALTO (Analyzed Layout and Text Object). 2004. Technical Metadata for Optical Character Recognition, version 1.2.

[2] See f.ex. http://openvault.wgbh.org/catalog/org.wgbh.mla:7376e451372c8a219648fc8e424aa9a1e8b463a4

mentioned above, and on new user requirements like plays in other manifestations, and social networking.

Generalisation of a book item is a book manifestation (item and manifestation concepts as defined in IFLA [3]). That means a manifestation in form of another edition, a translation, synthetic reading of encoded text, a live-recording of a play of a book containing drama, or it could be a manifestation in other dissemination formats like an e-book, a format for mobile devices etc.

Generalisation of citations is content segments, which can be an arbitrary part of the book, for example a chapter interval, a citation, a page interval, a literary work or the whole book. It must also support references that mark translated text, or references in connection with annotations, e.g. created by the public.

Other related author material can be anything from supplementary material to references to other dissemination platforms. Such material may also need to refer to parts of the material. For instance the sheet music may refer to a certain part of a play.

Social networking requirements are the most comprehensive generalisation of requirements. They are interesting for libraries, as a means to obtain corrections of digitisation, to get additional information on material, and to evolve interest groups as part of library life, for instance quizzes or student material related to the material [1]. Annotation may also come from research communities. An example is KB's involvement in the CLARIN project[1] which concerns infrastructure for scientific data. In CLARIN the ADL books are to be 'part of speech' encoded, where all words will be encoded with classifications of verbs, substantives etc.

General requirements will still apply, such as scalability, fast response time, user friendly interface. These requirements deserve special attention for a future context, since the magnitude and variation of data collections are increasing, which challenge scalability and fast response time. User interfaces should be homogeneous when they cover similar material digitised and represented in different ways. Search facilities set requirements for indexing and search in collections that may cover a range of material from many existing web applications.

An additional requirement comes from the growing demands for simultaneous display of different views and their interrelation. An example is synchronisation between audio and text e.g. using DAISY[2].

3.2 Requirements for Preservation

Our requirements for preservation are based on a decision to preserve digital born material and digitisation material to be reused in a future context. The preserved material will be the basis for a transformation into emerging dissemination and preservation formats. The assumption of reuse is the reason why we here only will consider a migration strategy. Emulation[3] does not support changes in presentation form and is therefore not considered.

From a preservation point of view, normalisation should be as simple as possible, and based as much as possible on standards in order to ease future understanding. Many different standards can support a final implementation. Examples are PREMIS[4] which provides a standard for preservation metadata, METS[5] which provides a standard to express object structure. Implicitly this also means that preserved data must not be structured in order to suit specific tools.

We need a flexible data structure for functional preservation in order to be able to represent a book object and its different migrations in the form of digital objects including structural and technical metadata. Furthermore, the relation between representations can become complex in the future, since we already know of cases where there are many-to-many relations between the digital objects, for example, many digital page images versus an e-book. A requirement is therefore to have a flexible object model where such representations and many-to-many relations can be modelled.

Another part of functional preservation is to preserve references into material, like citations references or future annotations. The modelling must therefore take into account how references into objects can be migrated as part of a full migration. The modelling must also allow creation of new versions with added contents as in the example of the missing pages.

Finally, it must be possible to store the data at differentiated confidentiality and bit safety levels, e.g. illustrations with copyrights have higher confidentiality than the rest, and the digital born material, such as author descriptions, needs a higher level of bit safety than the digitised book images, as long as the physical book is still available.

3.3 Interrelated Requirements

The interrelated requirements are the requirements derived from the interrelations between preservation and dissemination.

We will here view a logical object as a representation of an AIP (Archival Information Package) defined in the OAIS reference model. In OAIS, all preservation information is available in an AIP. However, not all information in an AIP is needed for dissemination. In OAIS the information for dissemination can be derived from enriched and transformed data.

We will require that logical object representations of the preserved data are relatively similar to representations in dissemination, and visa versa. The reason is that we will need to minimise processing time and storage cost for dissemination and preservation.

[1] Common Language Resources and Technology (CLARIN). http://www.clarin.eu/

[2] See http://en.wikipedia.org/wiki/DAISY_Digital_Talking_Book

[3] See e.g. "Keeping Emulation Environments Portable" (KEEP). http://www.keep-project.eu/

[4] PREMIS (Preservation Metadata Implementation Strategies). 2008. Data Dictionary for Preservation Metadata, version 2.0.

[5] METS (Metadata Encoding and Transmission Standard). 2009. Version 1.8.

When we focus on storage, we also need to analyse possibilities for reuse of stored data between dissemination platform and the preservation platform. For example, if they both use the same high consumption storage formats, they can share one copy used as part of the bit preservation. Sharing a copy should however be done with care [7].

Another possible cost-reducing architecture could be that dissemination relies on cache storage with a possibility to retrieve preserved data on request. In this case preserved data must be easy to identify and retrieve. However, also in this case the transformation from a preservation representation to a dissemination representation must be minimal in order to meet time and scalability requirements.

Note that these last requirements can mean an indirect requirement of coordinated shift in the preservation and dissemination formats. An example could be that dissemination of book pages was changed from TIFF to JPEG2000, and similarly for preservation.

4 DRAWING LINES TO THE FUTURE

In this section we will suggest a flexible object model for normalisation of data objects and their metadata, which can meet our requirements for functional preservation in a library context. Additionally we will point at possible implementations in a DOMS, on basis of current state of the art of library DOMS', and architectural and community requirements.

4.1 Suggested Shared Logical Object Model

This section will present a flexible object model which enables us to normalise the data in a way that respects our requirements for preservation and dissemination.

The suggested logical object model is meant as an abstract model which is respected in the explicit implementations. That means representations for dissemination do not need to be implemented in the same way as representations for preservation, although they do need to meet the requirement to retain a short route to dissemination.

The logical object model is inspired by an initial object model from the Planets project[1] and the additional work with a concrete implementation including simple ER-diagram developed in the Pindar project [6]. These object models support functional preservation including many-to-many migrations.

4.1.1 Representations

The logical object model operates with different object representations. A representation must be a self-contained representation of the object, independent of other representations. Examples are representations of different migrations, different versions, different derived versions etc. This is exemplified in Figure 2. The example given in Figure 2 could be a future version of ADL material, where page images have been migrated to JPEG2000, but the corresponding dissemination format is JPEG. Note that not all representations are preserved, e.g. the JPEG. Other examples of representations that could be added are synthetic voice or an e-book version.

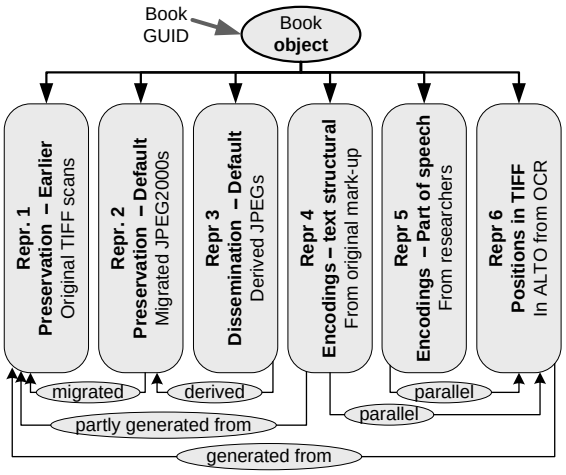

Figure 2. Example of representations of an object.

The different representations relate to each other in different ways. For example Repr. 6 was generated from Repr. 1 as part of the digitisation process. This is also the case for Repr. 5, but only partly, since it was enriched with manual encodings as well. For preservation and reproduction purposes, the technical details on how e.g. a representation is derived must be part of the metadata in the same way as technical metadata for a preservation migration, as e.g. described in the PREMIS standard.

It is not part of the model to define what kind of representations that can and must be included. It only prescribes that their relations must be described in detail. This creates a possibility for addition of new representations. It also creates a possibility to have more migration representations for one migration, which will be the case if different aspects of the original file will need to be represented in two different formats.

The concept of having representation also makes it possible to define groups of logical objects with common behaviours, both with regard to preservation aspects such as migration, and dissemination behaviours such as presentation in e.g. a web interface.

In the example, there are different encoded text representations. This illustrates a choice of keeping a split between different encoded texts for preservation, e.g. for positions, part of speech, and text structural encodings like chapters and stage directions in drama. This is especially preferable in a preservation perspective since the encodings are based on different parts of characters in the text, which will require encoding of overlapping hierarchies. This is a complex task, which contradicts the desire for simplicity in preserved data. Deriving and migrating information will therefore be harder, and there will be a risk of introducing errors in updates. Furthermore, positions

[1] Preservation and Long-term Access through NETworked Services (Planets). See http://www.planets-project.eu/

may deserve separate representation, since they only make sense for a very specific page image, e.g. separate position sets may come over time, and some may loose value due to deletion of related pages. On the other hand a disadvantage is that the OCR-text may have to be in all encoding representations. Note also that, even though some complexity can be eliminated by splitting up the encoding, there will be aspects where we cannot avoid some overlapping structure, as exemplified in [4].

In a future dissemination perspective where we want a dynamic environment, with frequent changes in the encoded text as a result of social networking, it will be better to have one source of update, i.e. a representation with all encodings including all overlapping trees, e.g. in an XML database. Such a representation could be added, as long as thorough description of relations to separated encoding representations is described.

The fact that dissemination is extended to include ingest operations in the form of quality checked corrective and extension information via social networking, complicates the interrelation between dissemination and preservation. Most preservation actions, e.g. bit preservation, can only be done on static material, thus the dynamic aspects will need to be represented in snapshots. The ingest process part must therefore be carefully considered, especially, if the encodings are represented differently. Furthermore, there will be a challenge in having asynchronous representations where the dissemination representation may be more correct than the preservation representation, as a consequence of social networking information that has not yet been quality checked and ingested.

4.1.2 Detailed Logical Object Model

A detailed logical object model must respect requirements for representation of many-to-many relations, referencing into objects, and a possibility to make corrections, e.g. by adding extra pages. Figure 3 illustrates the detailed logical object model by some book representation examples.

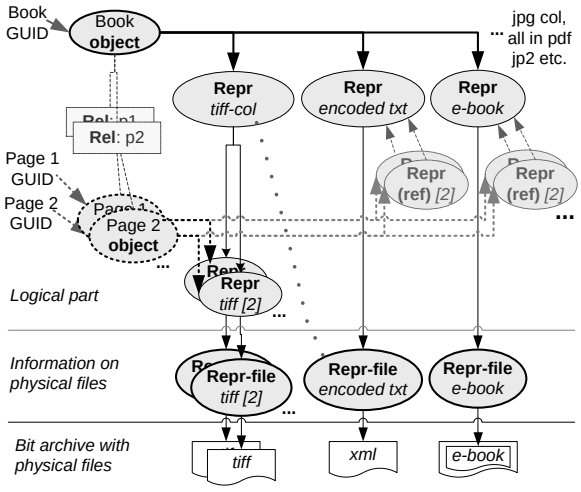

Figure 3. Modelling of a book object.

The broken lines and shapes in Figure 3 indicate that they are optional. The direction of arrows is not meant as a requirement of a concrete implementation, but an indication of the minimum information. This means, that a concrete implementation may have e.g. hasChild and isChild relations, although the arrow points one way in the model. The book object and the first layer of representations correspond to corresponding entities in Figure 2. In the *logical part* there are illustrated other levels of representations in form of pages. The part with *Information on physical files* is entities which make the link between the logical part and the physical files by referencing storage identification. The *Bit archive with physical files* is the storage, which possibly will be different according to the preservation level of the stored files. The dotted line from the *tiff-col* representation to the *xml* file for the encoded txt indicates that in the ADL case the order of elements in the collection is defined in the *xml* file. Note that the order representation could also be represented and preserved in a separate METS file, or it could be relation information metadata stored via a *Repr-file*. In the ADL case, this would mean that the current representation must be converted.

In the example there are page objects, which can relate to page representations for collection of TIFF files (*tiff-col*). However, there is no direct page representation for *encoded txt* or *e-book*, therefore, if desired, corresponding page representations in these book representations need to be made via references into the book representation. Note that such references may not make sense for all representations, e.g. the *e-book*.

It gets even more complicated when we want to model objects that represent a literary work. A literary work can be a poem, which starts mid-page x and ends somewhere in the first half of page y. Or a literary work can be a novel that spans over two volumes (book items). This means that a literary work can be defined at different levels in the logical part of the model.

Referencing into an object means addressing a part of an object. This reference mechanism should be transformable between different representations of the object, in order to ease the work of preserving the references in different preservation and dissemination forms, e.g. for migrations. References into objects are tricky. Normally, we would think of references based on atoms like a pixel in an image, a character in an ASCII text or time past in a soundtrack. However, a pixel may get another meaning in a migration. A character or its context may be changed due to corrections in the OCR of an encoded text. Furthermore in our example a pixel will have to refer to a page image in a book, which has a challenge related to the page numbering. Another challenge is that page numbers will not be part of e.g. an e-book representation, they will have a different meaning in a representation for a mobile device, and should have a different interpretation in e.g. a voice representation. If we consider encoding mechanisms e.g. using Xlink[1] this will again need consideration on how

[1] XML Linking Language (X-LINK). 2001. Version 1.0.

encoding is represented, updated and related to the different representations. Furthermore, in the ADL marginal notes example, the position reference of marginal notes was the most precise.

At a starting point, we will aim at a general reference mechanism which can be translated via relations between different representations, being aware that references like e.g. page numbers will not make sense in all representations. Similar referencing considerations will need to be taken for other formats such as sound, images and maps. In the future there will be an increasing demand for representations into objects, for example annotations added via social network communities. Such examples already exist, for example, for maps[1].

Part of referencing is also how we address objects or parts of objects with identifiers. Seen from a preservation perspective, identification of an object must be unique and persistent during time. Any semantics inserted into identifiers may confuse future uses such as e.g. a format extension or structure information which does not exist in the future. An example of a semantic free persistent identifier is Universally Unique Identifier (UUID)[2]. Identification of objects includes considerations on an object definition, in the sense that the object is addressable by the identifier in the future.

A choice must be made on how an object representation is identified in the future. For example, new versions of an object may occur in form of updates with added pages. Likewise for ongoing research reports, there may be several versions of a research report. The model does support creation of new versions, since adding of extra pages can be implemented by creation of a new *tiff-col* with a version relation to the existing *tiff-col*. Additionally a new representation would have to be created for related representation, e.g. for the *encoded txt*.

Many-to-many relations can be expressed in the model on the representation level, e.g. from a *tiff-col* to an *e-book*. When doing a many-to-many migration, the preservation metadata must include details of relations on the digital objects level. Many-to-many relations may also be needed in connection with reference translation between two representations, as described in the page reference example for Figure 3.

Annotations and information from social networking can be included in different ways depending on the type of information. Examples are; OCR corrections, part of speech annotations, relations to different material, or comments on author or text.

4.1.3 Consequences for ADL Data

As we have seen, the suggested logical object model can include special cases of the old ADL material, thus this data will be able to be reused. However, there will be a need for transformation of the data, which includes a risk of loosing data. Firstly, all page references must meet final identifier standards. Secondly, we may decide to have the structure of TIFF pages separate from the encoded text, for example in a METS file. A reason for this would be to have a less complex single representation of the preserved TIFF representation.

4.2 Possible Implementation

At KB we have reached the conclusion that community around preservation and dissemination is of great importance when deciding on the implementation of a DOMS. Another high priority is to have a system with high modularity and exchangeable components, where especially preservation issues must be system independent. Lastly, a high priority is to have a system with a homogenous treatment of similar materials.

There are many both national and university libraries that face the same challenges[3]. As this research also points out, we live in a time of rapidly changing demands for what a DOMS must cover. Not all problems can be solved at once, therefore there will be different priorities, e.g. due to different focus on different materials. Thus at present no system exists which can cover all the challenges to come in the next decade. There will however be a community that faces similar challenges, and has varying overlap of priorities for implementations.

Fedora commons[4] in particular has evolved into such a community, although there are different Fedora-based applications[5] like eSciDoc, Hydra, Islandora. Fedora has an advantage in being highly flexible with regard to how the data is modelled. A disadvantage, as well as a consequence of this flexibility, is that Fedora is far from a DOMS in the sense of being an off the shelf product. Furthermore, the Fedora-based applications are primarily focussed on dissemination aspects. Yet the Fedora case seems the best alternative to meet requirements of community and ability to model data in ways that complies with the logical object model.

The flexibility in Fedora opens many ways to make a solution that respects the logical object model, e.g. by using Fedora objects solely, or by encapsulating some of the modelling aspects in use of e.g. METS. This must however be done with care[6].

High modularity and exchangeable components are important for survival of the system, in which possibilities for renewal, enhancement and maintenance of the system are vital in order to meet new demands as a consequence of new technologies for formats and dissemination. The modularity requirement is also met by most of the Fedora initiatives. The Hydra initiative meets it, even to the extent that Fedora may be exchanged with a system offering similar functionality.

[1] Google maps, see http://maps.google.com
[2] UUID, see http://tools.ietf.org/html/rfc4122

[3] Several examples can be found e.g. in OR proceedings, for example the Mounting Books Project described on http://smartech.gatech.edu/handle/1853/28425
[4] http://www.fedora-commons.org/
[5] http://www.fedora-commons.org/confluence/display/FCR30/Getting+Started+with+Fedora#GettingStartedwithFedora-applications
[6] See e.g. OR 2009 contribution about Fedora 3.0 and METS on http://smartech.gatech.edu/handle/1853/28470

A system related requirement that of the possibility for different data to be stored under different confidentiality and bit safety levels. Although it is not part of Fedora, it is possible to implement this via workflows that handle insurance of storage in differentiated ways, and through implementation of access layer respecting confidentiality aspects.

The DOMS will end up as a system where ADL will be included as a special collection, possibly with separate web interface for ADL branding. Today there exist many different small applications like ADL, which all are part of dissemination from KB, but based on different frameworks. An example is www.tidsskrift.dk which disseminates digitised journal material produced with METAe[1] into a different format and using a different navigation than ADL. However, the cost of maintaining the different applications continues to increase. Therefore ADL and similar applications will be transformed into an integrated DOMS where preservation and dissemination aspects are treated jointly. This will be in line with requirements related to homogenous user interface for dissemination and ability to integrate with other systems.

5 DISCUSSION

The ADL case study represents relatively simple cases of material. We have argued that parallels can be drawn to other materials such as images and sound. There will, however, be other characteristics for other digital materials, which need to be investigated further.

There is still a challenge to settle on a general mechanism for proper referencing into objects. We may end up with different referencing mechanisms for different types of object representations. The selected mechanism must be taken into account in migrations, since inaccuracies in migrations can mean inaccuracies in migrated reference. In any case it may be hard to foresee the endurance of strategies for referencing.

Another related question is how to handle deletion of older versions or representation. Especially if references into objects rely on special representations (like positions) then the migration must include migration of similar referencing mechanism.

Having different representations of encodings in preservation and dissemination will add sources of error. This is a balance needing risk assessment and prioritising between meeting different requirements.

There are areas of the model that are not fully described as for example how to document relations between different representations. At this stage it is not necessary to make these processes and entities explicit, but they will have to be explicit in an implementation. As for any part of the data, the bit preservation level of the descriptions must be classified and effectuated.

Another area is versions contra representations. It is not a computer scientific question whether a new edition of a book is a new version with a new object identifier, or whether it is a new representation of an existing one.

[1] See http://meta-e.aib.uni-linz.ac.at/

6 CONCLUSION

We have argued that demands on preservation are closely related to demands on dissemination in a library context. Dissemination has many dynamic aspects and preservation tends to aim at static aspects, focus and goals differ, and thus demands on both preservation and dissemination will add complexity when viewed jointly.

We have presented a logical object model for normalised data that can meet the preservation requirement, including dissemination considerations and future requirements for new types of representation and information from social networking.

The hypothesis that we can use old digitised data in a normalised form will hold as long as the material is transformed, which is plausible, but does also involve risk of losing data.

The next step is to update the preservation strategy according to the findings, and to develop a DOMS for all digital materials in the library. This will include more thorough analysis of the challenge to reference into object and settle for a final implementation.

7 REFERENCES

[1] Holley, R. "Many Hands Make Light Work: Public Collaborative OCR Text Correction in Australian Historic Newspapers". *Technical Report from National Library of Australia*, Australia, 2009.

[2] Kejser, U.B. "Preservation copying of endangered historic negative collections" *Proceedings of the IS&T Archiving Conference*, Bern, Switzerland, 2008.

[3] Riva, P. "Functional requirements for bibliographic records: Introducing the Functional Requirements for Bibliographic Records and related IFLA developments" *Bulletin of the American Society for Information Science and Technology vol. 33 issue 6*, 2008.

[4] Sperberg-McQueen, C. M., Huitfeldt, C., "GODDAG: A Data Structure for Overlapping Hierarchies" *Lecture Notes in Computer Science, vol. 2023/2004*, Berlin, Germany, 2004.

[5] Zierau, E., Jensen, C. "Preservation of Digitised Books in a Library Context". *Proceedings of the International Conference on Preservation of Digital Objects*, Vienna, Austria, 2010.

[6] Zierau, E., Johansen, A.S., "Archive Design Based on Planets Inspired Logical Object Model". *Proceedings of European Conference on Research and Advanced Technology for Digital Libraries*, Aarhus, Denmark, 2008.

[7] Zierau, E., Kejser, U.B. "Cross Institutional Cooperation on a Shared Bit Repository". *Proceedings of the International Conference on Digital Libraries*, New Delhi, India, 2010.