# PRESERVATION OF METADATA

## *A case study of a strategy to ensure technology agnostic metadata preservation*

**Eld Zierau**
*Royal Danish Library*
*Denmark*
*elzi@kb.dk*
*ORCID: 0000-0003-3406-3555*

**Abstract – This paper focuses on strategies for preservation of metadata; one of the major topics in the creation of a new digital preservation strategy for the merged Royal Danish Library. Preservation of metadata is important to ensure preservation of all relevant information in order to be able to access digital objects in the future. This includes all kinds of metadata, which contributes to the understanding of an object, e.g. preservation metadata as well as descriptive metadata.**

**The need for metadata is commonly accepted, but it is not as commonly accepted that metadata need to be preserved to the same extent as files. There are many challenges due to the fact that metadata are often updated. This is probably one of the reasons why there exist numerous examples of metadata being under backup only and not under bit preservation.**

**Preservation of metadata is not just needed for future access of the objects, but also for re-establishment of repository systems after major breakdowns – or for establishment of a new repository as part of an exit strategy from a previous system. The latter case may also mean that the metadata have to be structured in a way, which can be used by different systems supporting digital preservation.**

**This paper describes how the newly merged Royal Danish Library has created a digital preservation strategy to cope with these challenges, and discusses the background for choices made in this process.**

**Keywords: metadata; preservation; data model; strategy; policy; exit strategy**

**Conference Topics: Designing and Delivering Sustainable Digital Preservation; The Cutting Edge: Technical Infrastructure and Implementation**

## I. INTRODUCTION

This paper describes how Royal Danish Library will ensure proper preservation of metadata. This is partly formulated in a new digital preservation policy (overall level) and strategy (more operational level) [1,2]. The policy and strategy reflect a merge of different preservation policies and strategies from two former libraries. Before 2018, Denmark had two national libraries which were each responsible for different types of materials, e.g. one library was responsible for collecting and preserving Danish Radio and TV production, while the other library was responsible for collecting and preserving Danish texts and computer games as well as donations from deceased authors. Furthermore, the two libraries were jointly responsible for the Danish web archive. In 2018, the two libraries were merged and named Royal Danish Library, and consequently three different digital preservation policies and strategies had to be merged and aligned in one set of digital preservation policy and strategy.

The purpose of this paper is to give an example of how this new strategy can align different digital preservation strategies and enable preservation of metadata along with exit strategies for different current and future systems. Additionally, the new strategy enables access to all metadata for all materials independent of the system from which they originate. Furthermore, it is our hope that this

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

DOI: 10.1145/nnnnnnn.nnnnnnn

iPRES 2019

paper can form a basis for feedback and further discussion of metadata preservation strategies.

Most organizations use some sort of system to support their digital preservation. However, many systems do not support preservation of metadata, but only preservation of files, when it comes to the low-level bit preservation. Examples are Preservica[1], Cumulus[2] and most (if not all) Fedora systems[3].

Even if systems do support some sort of bit preservation of metadata, the format of metadata usually depends on the system. Actually most systems have their own way of structuring and exporting metadata, therefore it is important to stress that the points made in the following description is *not* a critique of the above-mentioned systems, but examples of a general challenge. This is also why the systems are not addressed as preservation systems, since no current system covers all the needed digital preservation support. Examples of non-standardized formats for metadata in commercial preservation supporting systems are e.g. XIP in Preservica and DNX in Rosetta[4]. The same can be said for many locally developed systems, and systems built on open source software like Fedora. Examples are a Fedora 3 based system at Royal Danish Library in which the Fedora 3 FOXML was used as container for metadata; and another where at some point metadata were preserved from a Cumulus system with locally key/value defined metadata.

Different efforts have been made in order to deal with a standardized way to structure and pack metadata. For example, METS[5] is a standard for encoding descriptive, administrative, and structural metadata regarding objects within a digital library. Today, METS is used as a container format for wrapping different metadata, and several initiatives like e.g. E-ARK[6] have METS as part of their framework with an additional recommendation of how to get a more standardized way of representing metadata.

Even within the standardized ways to structure metadata, it is a well-known fact that there is no one standard to fit all, and there are many ways to combine the use of different standards, e.g. combination of MIX[7], METS and PREMIS[8] [3].

This paper describes the choices made by Royal Danish Library to fulfill the purpose of having independent metadata preservation. The descriptions are accompanied by the reasons for choices to enable a basis for discussion as well as reuse of argumentation for organizations with similar challenges.

The paper will start by providing a general description of the importance of metadata preservation along with the implied requirements for and challenges in metadata preservation. Next, the relevant strategies and policies for metadata preservation is described. To illustrate the strategy, the paper includes a detailed preservation example on several systems.

## II. METADATA PRESERVATION

Preservation of metadata is important to ensure preservation of all relevant information in order to be able to access digital objects in the future. This includes all kinds of metadata, which contribute to the understanding of an object, e.g. preservation metadata as well as descriptive metadata.

In the Open Archival Information Systems Reference Manual (OAIS), this is a question of preserving the full Archival Information Package (AIP) [4] p. 1-9:

> "An Information Package, consisting of the Content Information and the associated Preservation Description Information (PDI), which is preserved within an OAIS",

where Content Information is [4] p. 1-10:

> "A set of information that is the original target of preservation or that includes part or all of that information. …"

and where Preservation Description Information is [4] p. 1-14:

> "The information which is necessary for adequate preservation of the Content Information and which can be categorized as Provenance, Reference, Fixity, Context, and Access Rights Information."

[1] https://preservica.com/
[2] https://www.canto.com/cumulus/
[3] https://duraspace.org/fedora/
[4] https://knowledge.exlibrisgroup.com/Rosetta
[5] http://www.loc.gov/standards/mets/
[6] http://www.eark-project.com/
[7] http://www.loc.gov/standards/mix/
[8] http://www.loc.gov/standards/premis/

In other words, the AIP needs to include all relevant metadata for the package, not only for future access, but also in order to understand its context (not necessarily implemented as a package, but as information that is findable). Thus, metadata need to be preserved and bit preserved to the same extent as the data they describe.

Even at the bit preservation level, preserving metadata is not an easy task, since they are usually dynamic in the sense that they can be frequently updated. Following preservation guidelines strictly, all versions of metadata must be preserved with an audit trail containing information about what was changed and when. In practice, this may be a resource consuming task in cases where there are frequent minor corrections. A different strategy could therefore be to accept risk of loss of some information by pooling changes, and ingest them into preservation at longer time intervals and reduce the requirements of audit trail information. In any case, this would require some sort of version control of metadata.

To conduct preservation activities, all the preserved data and metadata must be available for (at least) digital preservation activities via some sort of repository system. Taking into account the rapid changes in technology during the last decades, long-term preservation will eventually imply the following requirements for a repository system: 1) repository software must be exchangeable 2) repository must be re-establishable from preserved data.

In other words, metadata must be preserved, and exit strategies for running repository systems must exist. These were also the main requirements taken into account when formulating the sections regarding metadata in the newly merged Royal Danish Library's Digital Preservation policy and strategy.

No matter which preservation supporting system a repository is currently using (commercial or non-commercial), there is a very high probability that the system will have to be replaced later on since this is about long-term preservation. This leaves some hard choices in case the preserved metadata are structured in a way that is dependent on the system, which must be replaced. One option is to let the metadata remain in the same system dependent format, which over time will result in different metadata structures from different systems. Another option is to convert the metadata. For large amounts of metadata, this can be a huge task, with the added risk of losing information during the conversion. The Cumulus based metadata in Denmark were converted, which resulted in a project lasting more than a year, even though the amount of data was relatively small[9].

## III. Independence of Preservation Supporting Systems

One of the major steps for the newly merged Library was to formulate a common policy and strategy for digital preservation covering all variations of digital materials.

The goal for both former libraries has always been to ***preserve data and metadata*** in a form that can be interpreted and understood in the future. Therefore, both former strategies focused on the use of appropriate ***standards*** for metadata, implementing a ***technology watch***, and basing all digital preservation decisions on proper ***risk management***. To reach this goal, both libraries aimed at becoming a ***Trustworthy Digital Repository***. This implies an aim to create a robust ***organizational*** anchoring of the work with digital preservation, in a way that maintains consciousness and responsibility of digital preservation as one of the key tasks for the libraries. Consequently, these policies could easily be mapped into a common policy for the merged Royal Danish Library.

However, the detailed strategies to achieve these goals differed a lot at the two former libraries. While the former State and University Library had to handle daily deliveries of huge amounts of homogenous data for TV transmissions, the former Royal Library had to handle heterogeneous materials with varying levels of confidentiality and complexity. Furthermore, the libraries had chosen different systems to support their preservation; just before the merger one of the libraries signed a contract with Preservica to replace the existing Fedora 3 based internally developed system, - and the other library relied on a Cumulus repository system with preservation services built around it.

The merged Royal Danish Library of course aims at creating a uniform ***Technical infrastructure*** with digital preservation processes to optimize as much as possible when taking into account the

[9] About 185000 records were converted.

requirements for bit safety, confidentiality, accessibility and the complexity of the digital materials. This means that the new strategy had to take into account an aim at as few systems as possible on the long run, and to at least get an as uniform way to handle preservation as possible. Since the market of supporting preservation systems is constantly moving, this has led to a strategy of keeping the Library as independent as possible of the systems used to support its digital preservation activities. This has resulted in a strategy of an intermediate infrastructure (as a goal for the next three to five years) illustrated in Fig. 1.

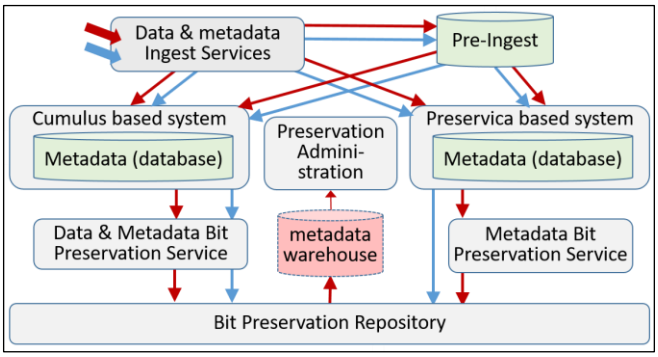


Figure 1 Strategic Royal Danish Library Technical Infrastructure.

The blue arrows specify flows of metadata, while the red arrows specify data flow.

In this infrastructure input data and metadata will be made as uniform as possible by common ingest services. Realistically, ingest of data and metadata cannot be processed by preservation supporting systems right away. The reasons for this are many, for example, the new types of material may need extra processing, implementations to the existing systems may need adjustment, and there are bulks of materials to be ingested, which must await capacity expansion. This is also why the architecture includes a pre-ingest area.

The number of supporting preservation systems will be reduced, since it is expected that all data preserved in the old Fedora 3 based system will be migrated to Preservica.

Metadata from Preservica will be bit preserved by a special metadata bit preservation application, which will extract metadata from Preservica and bit preserve them in the agreed standardized form, in the same way as the already existing metadata preservation application for Cumulus does.

Finally, a metadata warehouse is planned with an extract of the bit preserved metadata in a standardized format, which can provide an overview of all preserved metadata in a uniform way independent of the systems handling preservation. The metadata will be technology independent and can therefore cover metadata from all current systems, thus the warehouse can also be basis for support systems, e.g. preservation administration in the form of preservation planning.

## IV. Metadata Strategy

Royal Danish Library has formulated a vision for data and metadata in the new digital preservation strategy: "Royal Danish Library bit preserves data and metadata in a form, which enables the material to be understood and interpreted in the future".

Bit preservation of metadata is carried out to ensure preservation of materials that are to exist on the long term, but also to protect considerable investments in the form of time and resources spent on the creation of metadata for digitized materials with existing analogue copies.

The Library's strategy is to preserve in formats suited for digital preservation (both regarding data formats and metadata formats). To the furthest extent possible, the Library will use open formats, which are standardized or internationally acknowledged. Furthermore, the strategy is to use the metadata formats in a way as close to international standards and best practice as possible. The purpose of these choices is to increase the probability of understanding the metadata in the future and the example with Cumulus data showed that non-standardized metadata can become hard to interpret very quickly.

Royal Danish Library aims to preserve its digital collections in as few metadata formats as possible. However, a full normalization of metadata formats in preservation will never be possible. The reasons are that there are many specialized metadata formats for different types of data (e.g. MIX for still-images only) and many tools to support export of metadata in non-standardized formats exist. Furthermore, the two former libraries used different standards for descriptive metadata. The Royal Danish Library does not want to change existing metadata, since conversion of metadata from one format to another is a non-trivial task with risk of data loss.

In order to be able to interpret and understand data and metadata in the future, it is necessary to carry out a number of tasks in relation to how data and metadata are interrelated. This includes

employment of a data model for metadata, to enable an independence of technology in general.

A data model includes relations and identifiers to identify related items, therefore the strategy explicitly states that relations to data must be preserved and that the Library uses universal, unique and persistent identifiers for identification of the preserved digital materials.

When digital objects and/or metadata are updated or changed, the new versions must be bit preserved along with a log of the changes, i.e. an audit trail for both material and metadata. These audit trails are regarded as metadata as well, and should therefore be bit preserved.

It is an aim for the Library to use combinations of the same metadata standards in as similar a way as possible. This will streamline the general preservation and make it easier to develop general access platforms for the preserved materials.

Generally, Royal Danish Library wants to preserve the following types of metadata:

- *Descriptive metadata*
  with information describing the content of the digital object
- *Administrative metadata*
  with necessary information for curation of the digital object, including:
  - *Technical metadata*
    e.g. file format, checksum and digitization information. These metadata are obtained by characterization of the material or delivered from the digitization process. The characterization is performed as early as possible in the lifecycle of the material. The output from the characterization is preserved along with information about the tools performing the characterization
  - *Preservation metadata*
    which includes the necessary metadata in order to perform digital preservation actions, e.g. level of bit preservation and logical preservation strategy
  - *Digital provenance*
    which includes audit trails for actions performed on the digital object. This includes metadata from the creation of an object, e.g. the scanner's serial number for digitized material, speed of a record player used for digitization, or software which creates an iso-image from a hard drive
  - *Rights metadata*
    which can be used to deduce who can be granted access to the material
- *Structural metadata*
  with information about structures, which may be inherited in the digital object (e.g. references to pages in a book object)

Royal Danish Library structures its metadata as illustrated in Fig. 2 (inspired by what The National Library of Australia has described [5]).

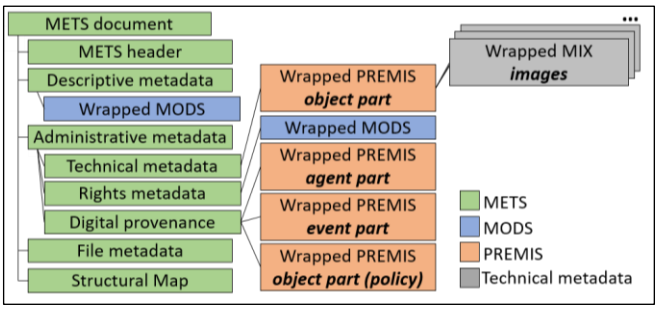


Figure 2 Metadata Model used at Royal Danish Library.

Some of the categories of metadata overlap, e.g. technical metadata are usually also regarded as preservation metadata etc. In practice this also means that metadata schemes overlap, and therefore can be used together in different ways, e.g. as described for METS and PREMIS in Ref. [3].

Royal Danish Library has chosen to use METS as container format for metadata, PREMIS for specific preservation metadata, and various XML based specialized standards like MIX for technical metadata.

In order to ensure long-term interpretation of metadata, the metadata profiles and schemes are publicly available at the website id.kb.dk, which is harvested and preserved by the Danish web archive. The name id.kb.dk is inspired by id.loc.gov for registries hosted at the Library of Congress.

## V. Data Model

The data model for bit preserved materials is designed for long-term use. It can contain data in a way, which supports delivery of whichever part of the data is needed for any application treating or publishing the material. In other words, any relation can be re-established by processing the bit preserved data. In order to ensure long-term interpretation of the model, it is designed to be as

simple as possible, without the optimization that front-end use scenarios may need. Such optimizations will be placed at other levels with additional information, which can support the optimization.

The data model is a simplification of the data model used in PREMIS and the data model used in the Planets project[10] (and partly used in Preservica).

This section will describe the data model on the theoretical and conceptual level, while the next section will describe a detailed simple example of how the data model is used in practice to preserve metadata at Royal Danish Library.

The data model has three basic data model entities (illustrated in Fig. 3 below):

***Digital Intellectual Entity,*** which expresses the top level of a digital object and unambiguously identifies a digital material. The object must be identifiable, regardless of which preservation actions, corrections or transformations have been carried out on the object through time. A Digital Intellectual Entity differs from a FRBR Intellectual Entity, since different manifestations of a FRBR Intellectual Entity would be interpreted as different Digital Intellectual Entities.

A Digital Intellectual Entity will always consist of one or more Representations, which represents versions of the Digital Intellectual Entity.

***Representation*** which expresses a Representation of a Digital Intellectual Entity, i.e. it represents a specific version of a particular Digital Intellectual Entity. A Representation can only represent one Digital Intellectual Entity.

The contents of a Representation can differ depending on whether it describes metadata in relation to a file or whether it describes metadata in relation to other Digital Intellectual Entities. Representations for a file usually contains metadata about the file and a reference to the preserved bit sequence representing the file. Representations for relations to one or more Digital Intellectual Entities contain metadata about the referred entities and their mutual relationship (if more than one). An example is the order of pages in a book, along with descriptive metadata about the book.

***File / Bit sequence***, which expresses a single file / bit sequence that is bit preserved.

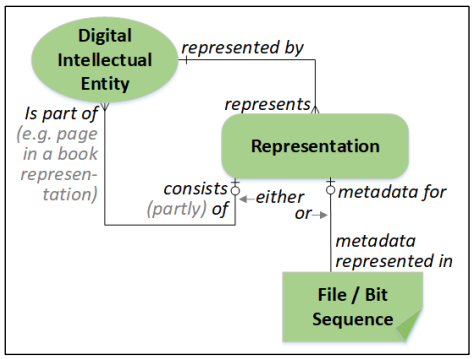


Figure 3 Data Model Entities and Relations.

There is no formula for how to model an object. Such decisions are made as part of the initial curation of the object, i.e. as preparation for or as part of ingest in a digital preservation supporting system.

## VI. A Detailed Postcard Example

The example illustrates how a digitization of a postcard is represented in terms of the data model and metadata scheme described above.


Figure 4 Digitized Postcard from "Majus smykker" of the H. C. Ørsted award 2013, photo by Grethe Aasted Therkelsen.

The postcard example (Fig. 4) is constructed to illustrate all possible changes. The postcard is digitized and updated in various ways over time by events in the following order:

a) Digitization of the front page of the postcard

b) Edits of metadata to the front page image (correcting Danish character encodings)

c) Re-digitization of the front page (because of errors in the first scanning)

d) Adding digitization of the back page (containing additional information)

### *A. Modelled Postcard Example*

The digitization steps are illustrated in Fig. 5 and explained in the following.

[10] https://planets-project.eu/

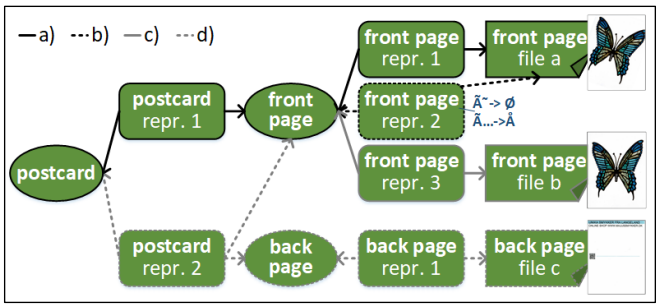


Figure 5 Data Model for steps in digitization of a postcard.

The digitization step a) produced a file and some metadata. In the data model, it produced the Digital Intellectual Entities and Representations along with the file. The Digital Intellectual Entity of the postcard only consists of an identifier, which is the valid reference for all versions (or rather Representations) of the postcard through time.

The Representation of the postcard "postcard (repr.1)" contains the relevant metadata for the postcard. This postcard Representation has no technical metadata, since no files are involved. However, it does contain "structural metadata" of the postcard pointing to the Digital Intellectual Entities of the front page. It cannot refer to the front page Representation, since this would mean that a minor page change would result in having to update the postcard Representation as well. This could easily start a chain reaction, since objects pointing to the postcard Representation would have to change as well. Consequently, it could become so large that it would require bit preservation of an additional large amount of data.

The front page Digital Intellectual Entity only consists of an identifier, which is the valid reference for all versions of the front page through time.

The Representation of the front page "front page repr. 1" contains the relevant metadata of the file. The result of the actual digitization is placed in the file "front page file a".

Step b) only consist of a metadata change, which does not affect the file itself, and thus the new Representation "front page repr. 2" is added with reference to the existing file as well as reference to the existing Digital Intellectual Entity it represents.

In step c) the "front page file a" is exchanged with "front page file b" containing a new digitization. This new file has new technical metadata, and therefore needs its own Representation "front page repr. 3", which refers to the new file and the existing Digital Intellectual Entity that it represents.

Step d) adds a back page to the postcard. A new Digital Intellectual Entity and Representation is created for the new back page in the same way as the first digitization of the front page. Since this is a change for the actual postcard, the change will also result in a new Representation for the postcard itself "postcard repr. 2", which includes the structural metadata of the postcard pointing to both the Digital Intellectual Entity of existing "front page" and the new "back page". Furthermore, it points to the Digital Intellectual Entity "postcard" which it represents.

There are many identifiers involved in this data model. The most important ones are the identifiers for the Digital Intellectual Entities, since these identifiers will be the reference point for digital material over time, in order to reference different versions (Representations) of the digital material.

### *B. Files Produced for Bit Preservation*

The metadata for the digitized files are represented in XML files with metadata that obeys the XML metadata schemes. Relations between the data model entities are specified by defining identifiers and specifying relations in the XML files. For example, Representation "front page repr. 1" refers to the "front page file a" through the structural metadata of the METS metadata (in METS files and METS structmap).

The reference to the Digital Intellectual Entity from the Representation is specified as part of the PREMIS metadata as a structural relationship (relationshipType: structural, relationshipSubType: represents and UUID for the Digital Intellectual Entity in metadata for Representation).

In practice, producing the XML files results in the creation of many small files. Even though technology has come far in relation to handling many small files, there are still issues. Therefore, the metadata are packed in chunks before being bit preserved. Royal Danish Library has chosen WARC for this purpose [6].

For optimization purposes, an extra WARC record is produced containing information about the relationship between Digital Intellectual Entities, each of their Representations and files (if a file exists for the Representation). This information is preserved along with the timestamp of archiving the Representation of the Digital Intellectual Entity. This extra information is redundant information, since it can be re-produced by reading all bit preserved metadata and finding the respective identifiers.

However, doing this would be a very time consuming process. An additional benefit is that WARC allows ”browsing” of the metadata, where the different versions of the Digital Intellectual Entity can be distinguished by the recorded timestamp, in the same way as browsing revisited web archived materials.

WARC packages with WARC records for each step of the postcard example are publicly available at id.kb.dk[11]. The actual bit preservation of the files is obtained by use of the bitrepository.org software [7].

## VII. METADATA PRESERVATION BASED ON EXPORTS

Royal Danish Library wants to ensure bit preservation of metadata through routinely export of metadata from the preservation supporting systems (i.e. Preservica and Cumulus systems), including audit trails for the individual digital objects. The exported metadata are structured as described in Section IV about metadata (illustrated in Fig. 2).

Based on the postcard example the following section describes how extracts of metadata from the existing systems are mapped into the technology independent general data model.

### *A. Postcard Exported from Cumulus*

This section describes how Royal Danish Library extracts metadata from the materials placed in a key/value based Cumulus system.

In Cumulus, the Library has defined keys for values used for transformation of the data into the general data model (except from identifiers and timestamps to be evaluated in the preservation process). Since Cumulus does not have any facilities for bit preservation, the Library has developed the program “Data & Metadata Bit Preservation Service” (depicted in Fig. 1). This application extracts files and metadata from Cumulus and transform the metadata into the structure described for metadata and the general data model, before sending it to bit preservation.

### *B. Postcard Exported from Preservica*

In this section, it is described how Royal Danish Library will extract metadata from the materials placed in Preservica.

The data model used in Preservica (hereafter referred to as the Preservica data model) has some resemblance to the data model described in this paper (called the general data model). Both data models are inspired by the data model from the EU Planets project, which ended in 2010. However, there are differences in terminology and meaning, especially regarding dynamic or static status of the Digital Intellectual Entity. Furthermore, the Preservica data model only takes into account the changes in files – not the changes in metadata. Resemblances and differences in concepts between the Preservica data model and the general data model are:

***Files*** in Preservica contain both a file *and* metadata for the file. Furthermore, these file metadata can be changed. To map Preservica’s data model to the general data model, one would have to separate the file from its metadata and have the metadata in a Representation for the file (and with changes to file metadata in new Representations).

***Manifestations*** in Preservica are similar to Representations in the general data model, (at least the preservation Manifestations)[12]. There are, however, two main differences. Firstly, metadata in Manifestations are changeable. Secondly, Manifestations in Preservica assume that there can be only one active preservation Manifestation at any given time. In the case of e.g. preserving a heavily animated PowerPoint presentation, there is a need for several different “preservation Manifestations”, e.g. a migration to a PDF version to preserve the look and feel of e.g. colors, and a newer version of PowerPoint to maintain the idea of the animation in the previous version. The general data model does not assume anything about whether Representations are active or not, and can therefore cover such cases.

***Deliverable Units*** in Preservica are similar to Digital Intellectual Entities in the general data model. Again, there is a difference in the fact that a

[11] The examples can be found at htttp://id.kb.dk/examples/

[12] Preservica also has presentation Manifestations, which are not concerned with the actual preservation and therefore not part of the general data model.

Deliverable Unit has changeable metadata. A Digital Intellectual Entity cannot change at all, since this would require a new identifier, consequently a Digital Intellectual Entity cannot contain metadata. In order to have metadata at this level, it must have a separate layer of Representations.

Preservica also has *Collections*, which are collections of Deliverable Units, and a Deliverable Unit can belong to one Collection only. If ignoring the latter extra restriction, there is no difference between a Collection of Deliverable Units and a Deliverable Unit consisting of Deliverable Units. Thus, Collections do not need additional comparison.

The above-mentioned postcard example will in the Preservica data model look as depicted in Fig. 6.

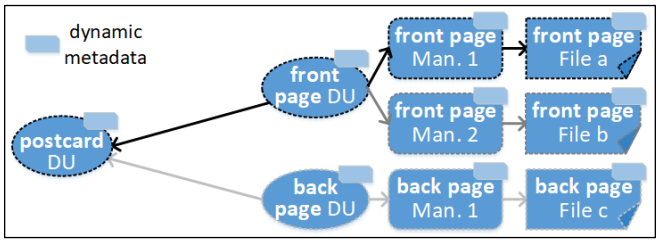


Figure 6 Postcard example implemented in Preservica.

The Preservica data model looks much simpler than the general data model, but this is caused by the fact that the Preservica data model does not support preservation of metadata and changes in these metadata for all Preservica data model entities (Collections, Deliverable Units, Manifestations and Files).

The difference from the general model is that changes are not represented. For instance in Preservica, the adding of the back page cannot be seen in the Manifestation of the postcard, since it is only represented as the back page file pointing to the postcard Representation. In the general model, this relation is explicitly recorded as part of the postcard Representation "postcard repr. 2". Furthermore, the change of metadata for the front page is not visible in the Preservica data model, since the Preservica "front page Man. 1" covers both "front page repr. 1" and "front page repr. 2" representing the metadata before and after the metadata update.

How to map the Preservica data model to the general data model is illustrated in detail in Fig. 7 for the front page (with a file).

In practice, there are two ways of mapping Deliverable Units with files to the general data model. Which method to choose depends on the type of metadata they cover, - or rather, how curators originally decided to place metadata on the Deliverable Unit and Manifestations in Preservica.

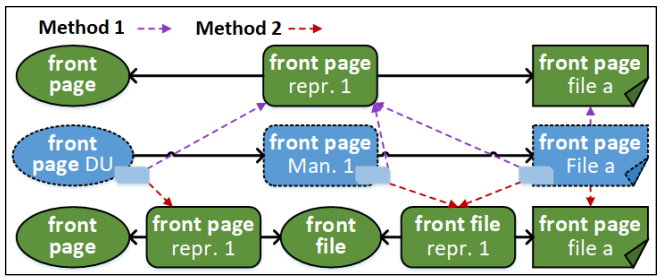


Figure 7 Two ways to map Preservica postcard (in blue) to general data model postcard (in green).

Technical metadata will always be on the File / Manifestation: the File and Manifestation are joint since the Preservica interface does not distinguish between whether metadata comes from a File or its Manifestation. This is also why all the Preservica metadata for the File and Manifestation (at a specific time) are mapped into the Representation of the file for both Method 1 and Method 2.

In most cases, the types of metadata placed on the File / Manifestation and the Deliverable Unit differ (are disjunct). Normally, descriptive metadata are placed on the Deliverable Unit in order not to repeat them for each Manifestation. Therefore, Method 1 is the most likely to be used.

Method 2 is used in cases where there is a need to distinguish between metadata for the File / Manifestation itself and metadata for the logical unit that the file represents (the Deliverable Unit), e.g. descriptive metadata. In such cases, an extra layer (the front file Representation and front file Digital Intellectual Entity) is inserted to distinguish between the two types of descriptive metadata.

The mapping of a Deliverable Unit *without* files (only pointing to other Deliverable Units) is simpler, since all metadata from the Deliverable Unit are mapped into a Representation of the corresponding Digital Intellectual Entity. However, in Preservica, the information about relations to other Deliverable Units comes from the Deliverable Units at a lower level. This means that the postcard Deliverable Unit has e.g. descriptive metadata only, while the structural information about the front and back page Deliverable Units' relation to their parent postcard Deliverable Unit can only be found in these underlying front and back page Deliverable Units. Furthermore, the history information about the

adding of the back page has to be found in the Preservica log. Since a Digital Intellectual Entity cannot have metadata, the metadata are placed on the Representations. Again, if there is a need to distinguish between changes e.g. in descriptive metadata and structural information, then an extra layer can be added.

Based on the described mappings, Royal Danish Library will develop a service to transform metadata from Preservica to the standardized metadata format which will then be bit preserved. We are aware that there are challenges to this transformation. One known challenge is calculation of the bit preservation level, which for instance needs information about the Preservica Storage adapter used for the data. Another known challenge is calculation of provenance metadata for tools used in Preservica. However, the challenges so far seem to be solvable, if the right APIs to extract metadata from Preservica are provided.

## VIII. Discussion

The way that the Royal Danish Library's digital preservation policy and strategy will ensure metadata preservation is by no means the only way to do it. It will always be a matter of considering which risks the organization is willing to take; number and nature of preservation supporting systems in the organization; and to which degree different materials need to be preserved.

First of all, the decision about transforming metadata before their bit preservation has an inherited risk of losing information during this transformation. On the other hand, a delayed transformation of system dependent metadata (when the system is eventually replaced with another) will also involve risks. The Library regards it as a greater risk to wait with the transformation.

Secondly, there are many ways to choose and structure different metadata, both in interrelations between the different formats and in the way interrelations between data are expressed in a data model. The choices described in this paper will therefore be debatable for other cases.

Concerning the mapping from Preservica's data model to the general data model, there are other cases than the included ones, e.g. cases where a Deliverable Unit consists of more files without an explicit Manifestation.

Finally, the concept of intellectual entities has been a confusing area for a long time, and it could be worth discussing whether the definition of a Digital Intellectual Entity can assist in the understanding of intellectual entities when we use the term in connection with digital assets.

## IX. Conclusion

This paper has provided a case study of one way to ensure proper preservation of metadata, as reflected in the new Royal Danish Library's digital preservation policy and strategy replacing three former and very different sets of policies and strategies.

The case study has included aspects of how to deal with preservation of dynamic metadata along with exit strategies for different current and future systems and access to all metadata for all materials independent of the system from which they originate.

The paper has also provided details of the decisions and reasoning made to ensure that implementation of metadata preservation can fulfil the different requirements to exit strategies, reestablishment after major breakdowns, and support of metadata warehousing.

We hope that this paper can form a basis for feedback and further discussion of metadata preservation strategies.

## References


[1] Royal Danish Library Digital Preservation Policy, archived in: netarkivet.dk, archival date: 2019-03-14T14:18:57Z, URI: http://id.kb.dk/documents/DB_Politik_20190109_UK.pdf [part], 2019.

[2] Det Kgl. Biblioteks digital bevaringsstrategi (only in Danish) in: netarkivet.dk, archival date: 2019-03-14T14:18:58Z, URI: http://id.kb.dk/documents/DB_Strategi_20190109.pdf [part], 2019.

[3] Digital Preservation Metadata for Practitioners - Implementing PREMIS, eds. Editors: Dappert, A., Guenther, R. S., Peyrard, S. ISBN 978-3-319-43761-3, 2016.

[4] ISO 14721:2012. "Space data and information transfer systems - Open archival information system (OAIS) – Reference model", 2012.

[5] J. Pearce, D. Pearson, M. Williams, S. Yeadon, "The Australian METS Profile – A Journey about Metadata", D-Lib Magazine, vol. 14, no. 3/4, March/April 2008.

[6] ISO 28500:2017, "Information and documentation -- WARC file format", 2017

[7] B. A. Jurik, A. B. Nielsen, E. Zierau, "Flexible Bit Preservation on a National Basis", *Proceedings of the IS&T Archiving 2012*, pp. 2-7, 2012.