

# Metadata for Presentation/Preservation of Physical Structure

Jörgen Nilsson

Lulea University of Technology, Sweden

*jorgen.nilsson@ltu.se*

**Abstract.** This paper investigates support for presentation of physical structure in existing metadata schemas used in digital preservation. Physical structure in this case is entirely focused on visual attributes for digital objects, in particular web-forms of the kind that is being used in many eGovernment services. In order to spot gaps and problem areas, practical examples in the form of demonstrator objects is constructed using PREMIS, and to a lesser extent utilising METS behavior section. METS is also tried as a wrapper for PREMIS. The new version of PREMIS (2.0) is found to be useful, much depending on the more refined support (both related to earlier PREMIS and METS) for significant properties.

## Introduction

There are several projects around the world that has been undertaken in order to preserve the physical structure<sup>1</sup> of digital objects. Some of those propose emulations, yet some utilise a kind of Universal Virtual Computer (UVC), and some suggest migration to a reliable format that retains the physical structure (e.g. Portable Document Format / Archiving a.k.a. PDF/A). All of those solutions have their own advantages and disadvantages, as have been discussed in several publications (Bearman, 1999; Granger, 2000; Holdsworth & Wheatley, 2001; Lorie, 2001; Lorie, 2002; Rothenberg, 1999a; Rothenberg, 1999b).

---

<sup>1</sup> Physical structure is Charles Dollars concept for layout, logotypes, colours, and other physical appearance (Dollar 2000).

Yet another way, which in some sense resembles the UVC, is to use "Multivalent" documents (Phelps & Watry, 2005). This technique involves describing the characteristics of documents in a way so that they can be interpreted by a special software (browser) that relies on descriptions of how specific types of documents should be interpreted. A pdf file is for example accessible through the browser without having a pdf-reader installed on the system. The software also supports adding annotations to the digital objects, and presentation of the documents through different "lenses" that for example zoom in on the document. The software is made in Java for portability purposes, and "behaviours" can be defined for different types of documents. Those behaviours are then saved in an XML structure.

An even more metadata driven approach is the one suggested in Metadata driven presentation of digital documents/records (Nilsson & Hägerfors, 2007a). This approach suggests that the presentation of records should be made with common tools and by using metadata to describe the physical structure of the digital documents. This method is intended to complement already existing metadata schemas with metadata attributes necessary for visual reproduction of a digital document. One of the cornerstones for this approach is the possibility to separate the data from the presentation, which means that the data still remains available for processing if desired, while the original (or at least intended) presentation also can be replicated.

Previous work (Nilsson & Hägerfors, 2007b) has shown that potential users value the visual representation of an digital document, while they at the same time also would like to have the content (data) available in a computable manner. To facilitate this – metadata descriptions of visual characteristics is used, while keeping the data separated and computable, instead of "flattening" the object into e.g. an PDF file.

## Method

Two metadata schemas intended for digital preservation (PREMIS and METS) were evaluated according to how they support the concept of preserving physical structure of the digital objects. This evaluation was done on a theoretical level, through document studies and interviews with system developers. The evaluation was based on if the metadata schemas had any direct support for metadata related to physical structure, or if they had indirect support by facilitating external (supplementary) metadata schemas. Based on findings in Nilsson & Hägerfors (2007b) and existing metadata schemas intended for digital preservation, demonstrator objects were developed. The demonstrator objects physical structure were described with metadata and fitted to the structures of PREMIS or METS. During the elaboration of demonstrator object, the metadata structures suitability for carrying visual attributes were analysed and gaps identified.

## Visual Attributes of textual documents

The work undertaken is chiefly related to digital documents of a textual character, which however may include pictures, logotypes and other visual formatting. Forms used in eGovernment are examples of such documents, that at least are considered for preservation (not everything is preserved), and it is this kind of documents that are in focus in this work.

As shown in previous work (Nilsson & Hägerfors, 2007b), users regard "machine processability" (e.g. the possibility to search directly after data, or make statistical queries over several documents) important and if they had to choose between machine processability and presentation of physical structure – they would almost exclusively choose machine processability. In the same research, the users did however point out the importance of visual attributes for interpretation, understanding, recognition and context of the material. This led to the conclusion that the users, in general, both wanted machine processability of the data, as well as support for an "original looking" presentation (although most preferred a "simplified" layout).

Both the processability and the visual appearance of the digital object can be considered as *significant properties*. Significant properties, although sometimes labelled differently, usually is described as "those components of a digital object deemed necessary for its long-term preservation" (Cedars, 2002a p. 15). This is a quite common view of significant properties (cf. Hedstrom & Lee, 2006; Knight, 2008; Wilson, 2007), and on a generic level, it is hard to be more specific about significant properties, since they, precisely as stated in the quote, will be considered per object, or at least per group of similar objects. Work has also been put into categorising significant properties and they are usually grouped (if at all) as suggested by Rothenberg & Bikson (1999) namely; content, structure, context, appearance and behaviour. Of these categories this paper is mainly focused on appearance since this is how the digital object is presented to the user, but content also comes into play since this is, exactly as it says, the content (i.e. the data).

Keeping the "data" available for processing, and describing the physical structure with metadata is the approach chosen to address the issue of "presentation and processability for long term preservation". In order to collect metadata about the physical structure, we have to look upon what attributes that need to be described. Typical visual attributes of a form would be:

- Fields            Placeholders for "data"
- Labels            Labels related to fields and images
- Images            Could be used for clarifications or emphasising of the content
- Logotype(s)      Images (usually) representing the publisher of the content
- Colours            Could be especially important on text, used for emphasising

- Font            The font face of the text in the form, can vary between elements
- Layout        The internal physical relation between the elements of the form

To what degree those attributes are deemed as necessary could of course differ from case to case, which also is the case with significant properties in large (Cedars, 2002a p. 15), and this list is mainly to be considered as a generic list.

### Usage of existing metadata schemas

It would be an understatement to say that usage of metadata in digital preservation is a novelty, it is not. And in line with the main ideas of this work, complementing existing metadata approaches, there are a couple that comes into play. Those metadata schemas are intended for general preservation metadata, which however is a rather wide term. Cedars have described preservation metadata as all of the various types of data that will allow the re-creation and interpretation of the structure and content of digital data that has been preserved (Cedars, 2002b p. 6), and in spirit of this, the metadata schemas dealt with here are quite large. The schemas are therefore only dealt with in the narrow focus of visual attributes and significant properties. We start of with the PREMIS Data Dictionary (PREMIS, 2008).

### PREMIS (PREservation Metadata Implementation Strategies)

Being an outcome of that many partners with previous experience from developing metadata schemas for preservation metadata, including Cedars (Cedars, 2002b), and OCLC (OCLC, 2002) decided to join forces, PREMIS do have extensive work put into it.

The data model might look simple in figure 1, but deeper down contains

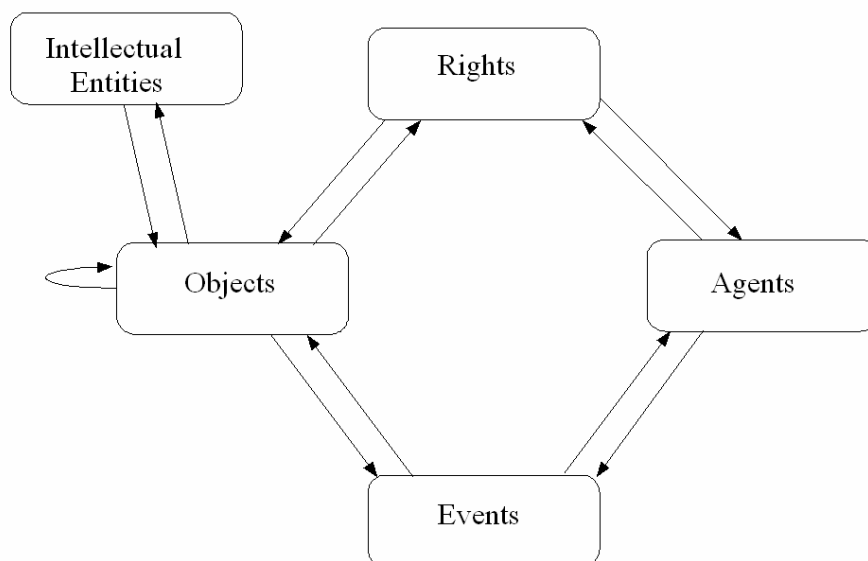


Figure 1: PREMIS Data Model (PREMIS, 2008)

several elements (95 for the Object entity, and 155 in all) deemed useful, far from all mandatory, for preservation of digital objects. In relation to this work, the most interesting part of the data model is Intellectual entities and Objects. The reason for this being that they are the parts that constitute the actual digital object, roughly it can be described as that intellectual entities are what we (humans) perceive as entities, and that those can be composed by several discrete digital objects. However, since Intellectual entities, consisting of one or several Objects, have no own metadata elements in PREMIS, we will only look upon the Object entity. Rights, events and agents are more related to the preservation activities and will not be dealt with here (PREMIS, 2008).

At time of writing, the PREMIS data model has just appeared in a new version (PREMIS, 2008) which has introduced some changes compared to the first version (PREMIS, 2005). Among the changes that are of most importance to us is that The metadata structure (figure 2) for an object have changed and *significant properties* has been moved out of object characteristics and are now on a higher level, and marked as extensible. Being marked as extensible is a new feature in PREMIS 2.0 and means that a section can include metadata encoded according to

```
1.1 objectIdentifier
    1.1.1 objectIdentifierType
    1.1.2 objectIdentifierValue
1.2 objectCategory
1.3 preservationLevel
    1.3.1 preservationLevelValue
    1.3.2 preservationLevelRole
    1.3.3 preservationLevelRationale
    1.3.4 preservationLevelDateAssigned
1.4 significantProperties
    1.4.1 significantPropertiesType
    1.4.2 significantPropertiesValue
    1.4.3 significantPropertiesExtension
1.5 objectCharacteristics
...
    1.5.7 objectCharacteristicsExtension
...
```

Figure 2: Excerpts of semantic units for an Object in PREMIS (PREMIS, 2008)

an external metadata schema, which was a bit of a hassle in the first version of PREMIS (Lavoie, 2008), in other words, precisely what we need. The metadata element (*semantic unit* in PREMIS lingo) designated for significant properties, aptly named *significantProperties*, has three sub-elements, which facilitates the use of key-value pairs (significantPropertiesType and significantPropertiesValue

respectively), or using an external metadata schema through significantProperties-Extension, for describing significant properties. This element is, for good or bad, quite loosely defined as "characteristics of a particular object subjectively determined to be important to maintain through preservation actions". Mentioned as example of a significant property was "content only" for a web page containing animations not considered essential to preserve, and "color" for a PDF with an embedded graph where the lines' colour were of importance (PREMIS, 2008). Object characteristics have also been given a new section called *Object characteristics extension* which gives room for "technical metadata" of a more specific character, primarily format-related metadata.

In PREMIS, the significant properties can be used for any type of object, while object characteristics cannot be used for the *representation* level. A representation is an object embodying an Intellectual Entity by combining a set of stored files and structural metadata (PREMIS, 2008), and the objects we are interested in are mainly of this character. Therefore the significant properties unit seem most appropriate for our purpose, whether using key-value pairs or using the extension feature.

#### METS (Metadata Encoding and Transmission Standard)

METS is a specification used for conveying metadata both for management of digital objects within a repository, and for exchange of such objects between repositories (or their users) (METS, 2007). A METS document has seven major sections:

- METS header Describes the METS document itself
- Descriptive metadata Contains or refers metadata used for finding and locating the object
- Administrative metadata Contains "typical archive metadata" such as provenance and IPR metadata.
- File section Lists all files that comprise the digital object
- Structural map The heart of METS, hierarchical structure linking metadata and files together
- Structural links Used to describe links between nodes in the hierarchy.
- Behavior used to associate executable behaviours with content in the METS object.

METS is quite flexible in allowing external metadata, and so PREMIS is known to be used for describing administrative metadata in METS. The file section has a convenient grouping feature, so that it is easier to group files related to a specific version together, if you for example have a scanned version of a several page document, the images can be grouped together, and another group

can be created for a text representation of the same object. This together with use of the *behavior* section, which is intended to link digital content with program code (or applications) in order to render the digital object (METS, 2007), can be useful for our purposes of displaying visual attributes.

## Development of demonstrator object

Main focus is on PREMIS 2.0 and its *significant properties extension* since it is explicitly defined as a part of the metadata schema. As a secondary case, use of the behaviour section in METS may be elaborated, but since its intention is to point to "executables", it may be to tightly coupled to software for my liking. Usage of METS as a "wrapper", utilising its *structural map* might still be a good addition to the PREMIS approach, and will probably be implemented regardless of whether the behavioural section of METS is adopted. The rest of this section describes the elaboration of a PREMIS-based demonstrator object.

Since the update of PREMIS, there are fewer gaps to be covered in these metadata schemas. PREMIS is however not intend to cover more specific significant properties, and this is where we come into play. The *significantPropertiesExtension* is explicitly pointed out for holding significant properties defined in an external schema, and it will be used for referencing the metadata needed for visual attributes. Some significant properties can be held as key-value pairs in the *significantProperties* section, but will be of a more generic character (e.g. *preserve look*). One drawback with using PREMIS 2.0 is that not much in the way of examples or other material is available for PREMIS 2.0 yet. Therefore there may be some implementation "errors" in the demonstrator object, but since the intention is not to fully follow the PREMIS specification as such, as much as using PREMIS as an example of a "wider" metadata schema that can carry significant properties specified for visual attributes, deviations are a nuisance but not critical for the demonstrator object intention at this stage.

During the development of the demonstrator object it became clear that a combination of PREMIS metadata and visual attributes metadata can lead to objects with a lot of metadata elements, though many of the PREMIS elements are not mandatory and could therefore be removed (in fact, there are only 4 mandatory semantic units at representation level in PREMIS) (PREMIS, 2008). The metadata is not intended to be entered manually, neither in PREMIS nor in our visual metadata schema, and therefore the amount of elements is mostly a problem from a complexity point of view, but it should not be neglected that complexity usually is a bad idea in long-term preservation of digital objects. One comment from a system developer concerned this, mentioning that PREMIS could be used in this way, but that it was unclear on how well it would work in practice. The demonstrator object was a tax return receipt and had the dimensions of approximately 800 by 1800 pixels. The number of metadata elements needed

to describe the visual appearance (i.e. layout) of the demonstrator object were approximately 73, and each of those elements had 5 basic attributes, namely id, x-coordinate, y-coordinate, type and description, which leads to ~365 elements. The layout was described in such a way that the textual data elements (including labels) were placed upon an image representing an empty tax receipt. Describing the entire layout (colours and lines and so on) would demand even more metadata and is considered, both by me and the system developer, as highly impractical.

## Conclusions

PREMIS and METS are both good foundations, covering much needed and necessary descriptive and administrative metadata. The update of PREMIS have closed some gaps that were apparent in the earlier version, and this is a good thing. At the same time (and for good reasons) the PREMIS group is trying to keep PREMIS quite "high-level", so they probably never will get close to the detail level that is needed for the approach suggested in this paper, which is not seen as a problem. Instead the extensibility of the schemas point at the importance of supporting metadata in an "onion" fashion, with several levels of metadata describing an object in different detail (and in different forms and shapes as well). Using those as containers for more specific metadata needed to facilitate reconstruction of the physical structure of forms, since they explicitly support external metadata for those purposes, would most likely be feasible. PREMIS is deemed as the most suitable for the simple reason that it has more explicit structure, and a specific unit for significant properties. This does not rule out the use or usefulness of METS, but for our purposes – PREMIS is more appropriate. The intention with this work was not to decide which metadata schema that was the best one, instead the focus lied on how metadata used to describe visual appearance could be used in conjunction with PREMIS or METS. One thing that quickly became clear is that the usage of the key-value pairs in PREMIS `significantProperties` is unsuitable for this kind of very detailed metadata, and that an external schema would be more suitable for this. The demonstrator object will be used (and improved) in further research involving digital curation experts.

## References

- Bearman, D. (1999). Reality and Chimeras in the Preservation of Electronic Records, D-Lib Magazine April 1999 vol.5 nr.4, ISSN 1082-9873, Available at:  
<http://www.dlib.org/dlib/april99/bearman/04bearman.html> (2004 03 13)
- Cedars (2002a). Cedars Guide to Digital Collection Management, Available at:  
<http://www.leeds.ac.uk/cedars/guideto/collmanagement/guidetocolman.pdf> (2008-04-08)
- Cedars (2002b). Cedars Guide to Preservation Metadata, Available at:  
<http://www.leeds.ac.uk/cedars/guideto/metadata/> (2008-04-08)

- Granger, S. (2000). Emulation as a Digital Preservation Strategy, D-Lib Magazine October 2000 vol.6 nr.10, ISSN 1082-9873, Available at:  
<http://www.dlib.org/dlib/october00/granger/10granger.html> (2004-03-10)
- Hedstrom, M. & Lee, C. (2002). Significant properties of digital objects: definitions, applications, implications. Proceedings of the DLM-forum 2002. Access and preservation of electronic information: Best practices. pp. 218-223. European Communities. Available at:  
[http://ec.europa.eu/comm/secretariat\\_general/edoc\\_management/dlm\\_forum/doc/dlm-proceed2002.pdf](http://ec.europa.eu/comm/secretariat_general/edoc_management/dlm_forum/doc/dlm-proceed2002.pdf). (2006-05-28)
- Holdsworth, D. & Wheatley, P. (2001). Emulation, Preservation, and Abstraction, RLG DigiNews vol. 5 nr. 4, (ISSN 1093-5371), Available at:  
<http://digitalarchive.oclc.org/da/ViewObjectMain.jsp?fileid=0000070511:000006283287&reqid=98851#feature2> (2008-04-08)
- Knight, G (2008). Framework for the definition of significant properties. InSPECT project Available at: <http://www.significantproperties.org.uk/documents/wp33-propertiesreport-v1.pdf> (2008-04-08)
- Lavoie, B(2008). PREMIS With a Fresh Coat of Paint, D-Lib Magazine May/June 2008 vol. 14 nr. 5/6, ISSN 1082-9873, Available at: <http://www.dlib.org/dlib/may08/lavoie/05lavoie.html> (2008-05-17)
- Lorie, R. (2001). Project on Preservation of Digital Data, RLG DigiNews vol. 5 nr. 3. Available at:  
<http://digitalarchive.oclc.org/da/ViewObjectMain.jsp?fileid=0000070511:000006279465&reqid=98851#feature2> (2008-04-08)
- Lorie, R. (2002). The UVC: a method for preserving digital documents – proof of concept. LTP report series nr 4, 2002.
- METS (2007). Metadata Encoding and Transmission Standard: Primer and Reference Manual. Digital Library Federation. Available at:  
<http://www.loc.gov/standards/mets/METS%20Documentation%20final%20070930%20msw.pdf> (2008-04-08)
- Nilsson, J. and Hägerfors, A. (2007a). Metadata Driven Presentation of Digital Documents/Records. In Stillman, L., & Johanson, G. (eds.): Constructing and Sharing Memory: Community Informatics, Identity and Empowerment (pp. 212-222). Cambridge Scholars Publishing.
- Nilsson, J. and Hägerfors, A. (2007b). Digital Archiving: Appraisal of visual attributes. Presented at Appraisal in the Digital World 15-16 november 2007, Rome, Italy.
- OCLC (2002). Digital Archive Metadata Elements. Online Computer Library Center, Dublin OH USA. Available at:  
[http://www.oclc.org/support/documentation/pdf/da\\_metadata\\_elements.pdf](http://www.oclc.org/support/documentation/pdf/da_metadata_elements.pdf) (2008-04-08)
- Phelps, T. & Watry, P. (2005). Proceedings of the 9th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2005). A No-Compromises Architecture for Digital Document Preservation.
- PREMIS (2005). Data Dictionary for Preservation Metadata: Final Report of the PREMIS Working Group. OCLC and RLG. Available at:  
<http://www.oclc.org/research/projects/pmwg/premis-final.pdf> (2008-04-05)
- PREMIS (2008). PREMIS Data Dictionary for Preservation Metadata version 2.0. PREMIS Maintenance Activity and Editorial Committee. Available at:  
<http://www.loc.gov/standards/premis/v2/premis-2-0.pdf> (2008-04-08)

- Rothenberg, J. (1999:1). *Avoiding Technological Quicksand: Finding a Viable Technical*  
Foundation for Digital Preservation, Council on Library and Information Resources,  
Washington DC, ISBN 1-887334-63-7. Available at:  
<http://www.clir.org/pubs/reports/rothenberg/pub77.pdf> (2004-03-13)
- Rothenberg, J. (1999:2). *Ensuring the Longevity of Digital Information*, Available at:  
<http://www.clir.org/pubs/archives/ensuring.pdf> (2004-11-01)
- Wilson, A. (2007). *Significant properties report*. InSPECT project. Available at:  
[http://www.significantproperties.org.uk/documents/wp22\\_significant\\_properties.pdf](http://www.significantproperties.org.uk/documents/wp22_significant_properties.pdf) (2008-04-08)