

Preservation Metadata and the OAIS model

A Metadata Framework to support the Preservation of Digital Objects

Report by OCLC/RLG working Group on Preservation Metadata (Published: June 2002)

Some comments

Hans Hofman¹

September 2002

Introduction

The reference models for preservation that exist at the moment, such as the well-known OAIS model, focus mainly on the functions and processes for preservation, not on metadata. That is however an essential part of the whole model. The Open Archival Information System (OAIS) includes a high-level, object oriented information model. So far some attempts have been made to produce sets of metadata for preservation purposes, but none covered as much of the OAIS reference model as the one now published by the Working Group on preservation metadata of OCLC/RLG.²

The Working Group has done a great job in trying to identify more precisely what metadata are necessary to preserve (certain types of) digital objects. Because of the predominance of the OAIS model any related contributions subsequently will have an impact on thinking on this issue. It is therefore important to be aware of the strengths and weaknesses or limitations of the presented metadata approach by putting it into a broader context. There are world wide more initiatives underway that try to deal with this issue. Examples are the ISO work on record management metadata, that also includes preservation metadata, and the work of the National Library of New Zealand, which is, though not yet published, another interesting and important development. And probably there will be others. Since the OAIS model in principle also could be applicable for electronic records, it will be interesting to see how this metadata report will or can accommodate records management and archival needs. Lastly questions arise, such as whether there should be more co-ordination, and whether there should be a (kind of) standard, and if so, how to achieve that and who and what disciplines should be involved then?

My comments are twofold, what can be said about this preservation metadata set in general and secondly to what extent does it cover the needs of the records management and archival community in preserving (archival) records? In doing so I will draw among others on the conclusions of the work being done in the Preservation Task Force of the first Inter Pares project regarding the preservation of authentic electronic records.³

1. The preservation metadata set in general

The terminology used by librarians is sometimes confusing for archivists, and the other way around I assume. Terms like provenance, archiving, context, records, etc. are used with slightly different meanings. This is already inherent to the OAIS model itself, so any discussion about preservation is challenged by confusion of terminology. It requires for records managers and archivists (and perhaps other information professionals) to be aware of these differences and to make a translation of such terms to their own domain.

The term 'records' is used for instance as a sub-sub-element of 'Resource Description' (under Reference Information) named 'existing records', which refers to existing metadata records. In the archival community this term has a very specific meaning, records document and are evidence of business activities. When applying the metadata model by a wider audience more awareness of the

¹ Hans Hofman is working as a senior advisor at the National Archives of the Netherlands and involved in several (national and international) projects in the area of digital preservation, such as the Inter Pares project and ERPANET. Furthermore he participates in the ISO TC46/SC11 records management committee, in particular as chair of the Working Group on records management metadata.

² Version of June 2002 available on www.oclc.org/research/pmwg.

³ The results are available on www.interpares.org.

issue of terminology is required, for instance by including clear definitions of key terms. The terms as used in the OAIS model and accordingly in this proposed set of metadata, such as digital objects, data objects, data content objects, information objects, have or may have a relationship with each other, but in other cases may not, or may overlap, or may have different meanings. The term digital object is the most ambiguous, since it refers to both conceptual and technical aspects, i.e. the intellectual content and its form, and the way it is digitally represented. That is at least the way I will be using it in this article.

In this respect it is essential to know the underlying concepts in being able to understand the metadata set and I think in this area more discussion is needed. It is, despite some definitions, not always quite clear what these concepts are or what they really mean.

Scope

The preservation metadata set as presented is a kind of compilation and elaboration of the work on metadata done in different projects (NEDLIB and CEDARS) and by different organisations (OCLC and NLA), and as such can be seen as very useful. Thus the background is mainly the library community, as is also obvious looking at the composition of the Working Group. The title suggests, however, that this proposed set of metadata is applicable to all digital objects in all domains. On p.3 it is indicated that it should 'accommodate the needs of the library community, along with other institutions tasked with the long-term management of information in digital form'. That is, however as I will indicate below, not the case and that is one of the restrictions to this proposed set.

In line with the OAIS model the point of view taken is that of the custodian, not being the creator. It supposes that at some moment in time digital objects have to be transferred, brought in, harvested, or ingested into the organisation and system the custodian is managing. The objects are preceding to this ingest under control of other organisations, like publishers or record creating organisations. The extent to which the management of objects can be influenced differs with respect to the type of objects. In the case of (government) records legislation is governing their creation and management, in the case of publications that influence will be mostly based on agreements between producers, publishers, and preservers. That includes also the way metadata are created and maintained. Ideally there should be a continuous and consistent creation and management of the metadata that accompany the digital objects, in practice that is still a lot to achieve in this respect. So, although the suggestion sometimes may be otherwise, preservation metadata does not only apply to what is under custody of a cultural or other preserving institution, but should be applied to the whole life cycle of digital objects. This implies life cycle management. I use the term of life cycle here as a chain of stages, such as conception or design, creation, maintenance and final disposition (as in the case of government records, destruction or 'sentenced' for having continuous value). During that life cycle the use of digital (and other) objects may change over time or across domains, because of different contexts, and that will have an impact on the creation and management of metadata. Preservation can be viewed as part of maintenance.

'Preservation metadata' is seen as information necessary to support preservation processes. In the report, however no link is made to these processes. An analysis of the processes and what information is needed to perform them would have been very useful to understand better this proposed metadata set. Without such an analysis it seems to me more difficult to identify the metadata needed. The Working Group may have considered it implicit since the set is based on the OAIS information model that again is derived from the functional model. The high level of that model however does not make clear immediately the required metadata. In this respect it is also essential to understand the nature of the digital objects that are to be preserved. There is an explicit statement that 'no assumptions about type or structure of digital resource' are made and that no particular preservation strategies are taken into account (p.3). But by taking library community needs as leading (be it implicitly) the approach is already restricting types of digital objects. Managing different types of 'digital objects', e.g. publications and records, may require not completely similar sets of metadata. Another issue is what requirements are governing the preservation processes? Approaching it from the object itself, as is done in this report is not enough. There needs to be insight and as a consequence also metadata about the preservation strategies, policies, and methods, together the context in which the preservation takes place. Although the OAIS model contains functions as 'preservation planning' and 'administration' these are not addressed in the report, may be because

they are not (yet) reflected in the information model. For a complete overview of preservation metadata nonetheless all functions have to be included.

Digital objects

Crucial question in understanding (long-term) preservation of digital objects is: what is a digital object? As already indicated that seems not as clear cut as one would expect. In the OAIS model an object of preservation is identified as a data object that can be either physical or digital. A digital (data) object is defined here as 'an object composed of a *set of bit sequences*', so it has to be seen as a (data) file that is stored on the disk and not as the 'thing' that is represented on the screen. The data object is accompanied by so-called '*representation information*' which documents the way the data object has to be interpreted in order to present it and to enable intended users to view and understand it. Together they form the information object. It is a rather technical approach that is taken here. Perhaps the real question should be, **what** do we want to preserve? Is it the intellectual content with the functionality it has to have in order to make sense and achieve its purpose, or is it the digital components that are necessary to reproduce it or both? One of the basic notions in a digital environment is the difference between what is shown on the screen and that what is stored on the disk or medium. There isn't even always or does not have to be a 1:1-relationship between them. A digital component may contain one or more records, or the other way around one record may consist of more than one digital component (e.g. in the case of a multimedia record). In many publications this distinction is not explicitly made and it is therefore not always clear what the subject of discussion or the object of preservation is. In general there is recognition of the disappearance of physical entities, but it seems as if the consequences of this notion are not always drawn. It is the message or the intellectual content which the author or creator intended to convey that has to be preserved. That content has a context, form, and structure and in some cases also behaviour (e.g. spreadsheets). Be it as it is, the terminology used is confusing and not consistent on this point. My view is that 'digital objects' should be seen as objects having both conceptual and technical aspects that are closely interrelated. As a consequence of the explanation given above a digital object may consist of more than one 'digital component'. The definition given in the OAIS model is therefore insufficient. Moreover, the implications of this concept for the model are not yet adequately considered. The distinction between digital components and intellectual objects as different views or entities, requires, as indicated above not a one-to-one, but a many-to-many relationship.

What does the OAIS information model say about the conceptual or intellectual aspects? Taking a closer look at what is meant with representation information it turns out to consist of two categories, '*content data object description*', and '*environment description*'. Each of these contain (proposed) elements that can tell something about the conceptual aspects of the object we are preserving, e.g. in the case of the first:

- 'significant properties', defined as 'properties of the Content Data Object's rendered content which must be preserved or maintained during successive cycles of preservation process';
- 'functionality', defined as '...any functional or "look and feel" attributes of the rendered Content Data Object, in regard to its current manifestation in the archival store; this intends to describe the current technical properties;
- 'description of rendered content', defined as '...Content Data Object's content, in regard to how it should be viewed and interpreted by users. Includes clarification of potentially ambiguous data, definition and description of data structures, etc.;
- 'documentation', defined as supporting documentation necessary/useful for display and/or interpretation of the Content Data Object
- in the case of the second the 'output format' of the '*Display/Access Application*', defined as description of the output to be expected from the Display/Access application.

It means we have no less than 5 metadata elements that could contain information on what should be rendered and presented on the screen. How all these elements relate to each other, if any, is unclear. Another issue in the approach here may be that the digital object is only considered to be a technical entity, and as such always seems to be at the centre of attention, not the intellectual object. Based on the question what should be preserved one would expect, that it would be the other way around, but it is not. To be more precise, in general we want to preserve the intellectual expression (object if you like) not per se the digital components of which it consists, because if one thing is obvious in a digital world, it is the fact that these digital components will change over time. Even rigid standardisation will not prevent that, although in the future adequate technologies may probably emerge that make things

easier in the area of preservation. That is also why in some reports the preservation of digital objects (i.e. their components) in their original format(s) is recommended. What we want to achieve though is that in the future we still will be able to see, read, and understand the documents or other information entities that were once produced for a certain purpose and in a certain context. In trying to achieve that we need to preserve these digital components of course, but as information technology will evolve, these components have to be migrated or in some cases emulated to be usable on future hard- and software platforms.

So one of the issues is to identify what the intellectual aspects are. The emerging notion of 'Significant Properties' seems to acknowledge that. In this metadata set that is identified as an element and may serve as description of intellectual aspects, if only it is established as such.

Finally special attention has to be paid to the so-called '*underlying abstract form*' (sub-element of '*Content Data Object Description*'), a notion that has been introduced and coined by the CEDARS project. Its definition says: a 'human readable description of the Underlying Abstract Form of the Content Data Object'. It intends to provide information about implicit structures (e.g. files and relationships) that should be represented correctly to render or access the 'Object'. As such it seems to address the relationships between different (digital) objects and not the structure of one single intellectual object for instance. To avoid this confusion I would like to suggest to make a clearer distinction between intellectual and 'technical' or physical aspects of digital objects.⁴

Administrative metadata

Another area that requires attention is the technical information needed to preserve and reproduce the 'digital objects'. That can be seen as part of administrative metadata. The section about 'Content Information' contains, apart from the 'Data Content Object', the category 'Representation Information'. Part of this category is 'Environment Description' that should provide information about the technical (hard- and software) environment, necessary to render the data object. It includes information about rendering programs, operating systems, computational resources, storage and peripherals. Information about the digital data object includes aspects as technical infrastructure of complex object, installation requirements, file description, so-called quirks (documenting any loss in functionality or change of look-and-feel), structural type, but also significant properties, etc.

Based on the given description of this section it only concerns the current technical environment, not information about any previous environments. Is the underlying idea that it will be sufficient to have information about the performed preservation activities such as migration, conversion and their results? That could be an approach, but it would have been useful to know the assumptions behind it. Another question that may be asked here is, whether also not information about the original technical environment should be kept or is this supposed to be part of the 'Content Data Object Description'? An understanding of the original technical environment, in which the digital objects were created, will help to preserve them. Of course, if the digital objects will be preserved in their original format, as in some cases is recommended, that information will be kept under the category discussed here. Nonetheless I would like to suggest including an element that reflects the original technical environment.

Most administrative metadata is subordinated under 'Provenance Information' (as part of Preservation Description Information) This category is meant to document the 'Object' as a dynamic entity, considering it as the result of a never-ending range of activities or an evolutionary process, without which it would not exist. The 'events' metadata are related to the processes or activities carried out in preserving data objects and should provide information about the management history. An issue here may be how to match that information to the requirements that are valid for the system? There has to be an evaluation process (in the OAIS model under 'Administration' Process) that takes care of that and will produce evaluation information (another set of metadata) that can be used to adjust (the performance of) the preservation function or system as such. These aspects are not discussed or included in the metadata set provided here.

⁴ In a recently published article Ken Thibodeau introduces even a third category, i.e. logical objects. It concerns how information is encoded in bits and the grammar (rules) that allows application software to interpret the data. As such it refers to e.g. the possibility of encoding the same conceptual object in different formats. See 'The State of Digital Preservation: An International Perspective. Conference Proceedings', July 2002; www.clir.org/pubs/reports/pub107/thibodeau.html.

2. The metadata set and preserving (archival) records

Records and records requirements

The second perspective I like to take in this article is that of the records and archival communities. The question is then, to what extent can the preservation metadata set fulfil the needs of preserving records? In order to be able to do so, it is necessary to have an understanding of what records are and what their requirements are.

Records, according to the recently published ISO records management standard 15489, is 'information created, received and maintained as evidence and information by an organisation or person, in pursuance of legal obligations or in the transaction of business'. So in order to understand, use, and interpret records correctly it is necessary to know their administrative or business context, as well as their interrelationship with other records created in the same context. In order to achieve that records have to be authentic, i.e. in short, they are what they purport they are. The main requirements for records to serve as evidence or authoritative information sources are therefore authenticity and integrity, and knowledge about the business context and about the interrelationship between records (e.g. in a case file). I will discuss them in the following paragraphs in relation to the metadata set further.

Authenticity: management

Surprisingly enough the issue of authenticity is hardly touched upon in this report. Does that mean that it is implicit or that it is not really seen as an issue? Unfortunately no explanation is offered. Since authenticity is one of the main requirements governing preservation of 'digital objects', publications, websites, or (archival) records alike, it will affect also the metadata requirements. Authenticity refers to the requirement to be able to retrace documents or records to their creation (or origin), so it will be possible to identify why, when, where, by whom and so on, they were created (or received) and used. Answers to these questions are needed not only to know the identity of a record, but also to know whether the presented information is trustworthy or reliable. In other words it should be possible to position a record in the time and the context from which it purports to originate. After all, if we know somebody or something (or think we know) we are able to establish whether we can trust him, her or it, or not. In order to enable this judgement or assessment we need information that can answer those questions. Especially in a digital environment authenticity has become an issue, because digital documents or records by their very nature are intangible and volatile, and easy to tamper with. Apart from information about their origin information about the management of the records or digital objects is necessary to be able to assess what happened since they were captured and whether something may have happened that has affected them in a negative sense. Finally the (conceptual) object itself has to be described, what are its essential characteristics? To some extent these categories of information may be found in the proposed metadata set. So can be indicated for example under 'context information' why an object has been created, be it in a rather technical sense, and under 'provenance information' what 'events' have taken place (management history), while under Content Data Object Description 'significant properties' are included, identifying the characteristics of the object that should be preserved. The question is, does that really meet the requirements for maintaining authenticity? It would have been helpful, if there had been more acknowledgement of the issue of authenticity and the requirements for it, and if the Working Group would have provided some background information about its view and considerations on this aspect and to what extent it is included or not.

Context

As indicated context information in an archival sense means information about the context in which records are created, i.e. the business activity and the responsible organisation. In this metadata set this is represented by the sub-category (of Context Information) 'Reason for creation', defined as 'documents information about why a content data object was created'. The accompanying explanation indicates that it regards mainly why a physical object was created. It limits the scope of contextual information to 'informational requirements associated with *managing the preservation process*'. As such it only refers to the role a certain data file plays, e.g. master file or so, a rather technical approach.

For information that helps to understand the background of the digital object is referred to 'Representation Information' (i.e. the section Content Data Object Description). In this part the element 'Documentation' is meant to provide documentation necessary to interpret the 'Content Data Object' and is assumed to be a link to where the documentation is (e.g. a URL). Perhaps this may represent the kind of information I mentioned above. Apparently this information is not considered as preservation metadata. Nonetheless it is essential for understanding the intellectual objects (publications, documents, or records) as represented on the screen and for being able to identify their authenticity. Furthermore that information has to be inextricably linked to the (intellectual) objects themselves. As such it has to be preserved as well and may be an object in itself. And in the case of records that kind of information or metadata will be created electronically and has to be preserved as such. In the Information Model and the derived preservation metadata set this part is not adequately addressed. In order to be able to preserve (archival) records it will therefore be necessary to extend the information model with another class of information that refers to business context. Such a subset could provide a structure for describing what in archival terminology is called information about 'provenance' (with a different meaning as in OAIS).

Relationships and Aggregations

An area that is not very much developed in this metadata set is that about relationships between objects. I already discussed the difference between the digital or better physical objects (i.e. a data file stored on a medium) and the conceptual objects (i.e. a publication or archival record as presented on a screen), which is not necessarily one-to-one. That conclusion entails more complexity and has its implications for defining metadata. Relationships in the proposed metadata set are now identified as a subset of Context Information (one of the sub-categories of Preservation Description Information). They consist of two sub-categories: 'Manifestation' and 'Intellectual Content'. These relationships can refer either to other *manifestations* of the same object (= the same content) or, in the case of Intellectual Content, to *relationships* between this and other objects (with related content). As such the relationships seem to refer to physical data objects. They cover only part of the relationships necessary to represent the complex network of physical and conceptual objects. Since an intellectual object, either a record or set of records or a publication, can consist of several digital components and the other way around, as discussed above, this area has to be enhanced.

In order to accommodate the identified complexity it is necessary to distinguish at least between the following categories of relationships:

- relationships between intellectual objects: for instance components of one collection, or in the case of records the components of a case file or a (archival) *fonds* or a series (representing different aggregation levels) and their position within that aggregation. In archival context this is called 'documentary context';
- relationships between the (structural) components of one intellectual object, e.g. the pages of a book, the elements of a record, the components of a multimedia document;
- relationships between digital components, e.g. those components that contain elements of one (or more) intellectual object(s); although this will implicitly include relationship between the digital components and the intellectual object of which they contain parts, this relationship should be explicitly described.

The first two categories will be fixed relationships, because they are inherent to the nature of the intellectual object (and can be considered as an example of significant properties). The relationships of the latter category will change over time, because technology will evolve and as a consequence the digital components will change.

Retention

An interesting issue for archivists is 'Archival Retention'. Unfortunately it is only touched upon rather superficially (p.42) within the area of the Ingest Process and it seems to refer not specifically to (archival) records, but to all kinds of objects. Confusing terminology again.

Although one may look at a repository based on the OAIS model as one that contains information objects that should be preserved forever, one may also consider the possibility of applying the model for more dynamic environments, such as record creating organisations. In that case the issue of appraisal and disposition of records has to be included. In that case the recently published records

management standard (ISO 15489) may serve as a useful framework. It would make the OAIS model even broader applicable.

Finally

Summarising the above comments I think the proposed set of preservation metadata provides an important building block in the search for preservation strategies. It is a first attempt to elaborate on the OAIS information model by synthesising existing results of different digital preservation projects in this area. There are some issues, however, which need further attention. They concern on the one hand the scope and the underlying concepts of the OAIS model and the resulting metadata set as presented and on the other hand the application of the model and metadata set in an records and archival environment. One issue is the scope of the metadata set, as it covers only metadata for information packages. It does not deal with other important areas of the OAIS Information Model as for instance Preservation Planning and Administration. It would be useful for a fuller understanding of preservation metadata, if that could be included in a next version.

The metadata set (and underlying model) would also benefit from a clearer understanding of the concept of authenticity and the way it is included. In this respect more discussion is needed on the (underlying) concepts, not only on how authenticity is seen and should be positioned, but also on what is meant with a 'digital object'. Especially the distinction between physical and conceptual or intellectual aspects of a digital object should be made more explicit and it will probably have an impact on the model and metadata set too.

More attention is also needed for the relationship between the (preservation) processes and the metadata. It will help to clarify the purpose(s) for which metadata is necessary. In the case of this report, however it is not clear how that is established or identified, or put otherwise, in what (preservation) processes what metadata is used. Such information would not only help to better understand the metadata set, but also to support a more effective use both of the OAIS model and the accompanying metadata model for preserving digital material.

Another point is the fact that the Working Group has taken the OAIS conceptual information model as the basis for its work. Since that model is a high level object-oriented model derived from the functionality of the OAIS model itself, the Working Group by definition had to stay on a high level as well in order to establish a set of preservation metadata. It has to be adapted and made specific to identified environments in order to be applicable. Nonetheless the proposed preservation metadata set seems to be more suitable for the library community than for other communities. In this article I have tried to articulate some of the main requirements for the records and archival community in preserving (archival) records. Based on that the conclusion has to be that some adaptations to the model and metadata set would be necessary to meet these. That regards such requirements as the concept of authenticity of records, information on the business context of records and on relationships between records ('documentary context'). In assessing the needs of the records and archival community the ISO records management standard 15489 may serve as a very useful framework. Such an exercise would also include a test for applicability of the model and metadata set for record creating organisations and as such broaden the view of the OAIS model. In a records environment preservation starts already in the design stage and requires a comprehensive life cycle management, not limited to an archival institution, but including all stages of the records life cycle. It again supports the suggestion to have a metadata set for the full OAIS model, that includes the management aspects of it.