The role of preservation metadata within the KB's preservation policy

Authors: Susanne van den Eijkel and Daniel Steinmeier
Date: 17 August 2021

With special thanks to: Inge Hofsink and Sam Alloing

Introduction

The goal of this document is to explain the role of preservation metadata within the National Library of the Netherlands' (KB) Preservation Policy.[1] This document is based on the Open Archival Information System (OAIS)[2] model, the guidelines for Trustworthy Digital Repositories (ISO-16363) and the metadata standard PREMIS (Preservation Metadata: Implementation Strategies). First, we will describe the background of this document and clarify what we mean by preservation metadata. Finally, we will focus on the role of metadata in relation to long-term accessibility of digital material.

In 2021, the KB intends to start a migration process to transfer their digital collection to a new preservation system. This will mark a new step in the history of storing digital material at the KB. The new system also provides opportunities for preservation. After focusing on bit preservation for the past 18 years, the KB is now considering functional preservation. This new approach underlines the importance of documenting the role of metadata and how it fits within the KB's Preservation Plan. Daniel Steinmeier (Digital Preservation Officer) and Susanne van den Eijkel (Metadata specialist digital preservation) have analysed the collections of the KB from the perspective of migration. The documents produced during these analyses focus on the choices that were made regarding the implementation of metadata. However, a specific policy for preservation metadata and how it supports digital preservation was missing up till now. The current document formulates some tactical principles in line with the KB's Preservation policy.

Within the KB, metadata used to be divided into two categories: Bibliographic metadata and preservation metadata. In this document we use the term preservation metadata in a broad sense. This means we focus on all the metadata that is necessary to support preservation goals, in order to keep digital material accessible for the long term. Preservation metadata is a subset of all the information that is needed to make preservation possible.[3]

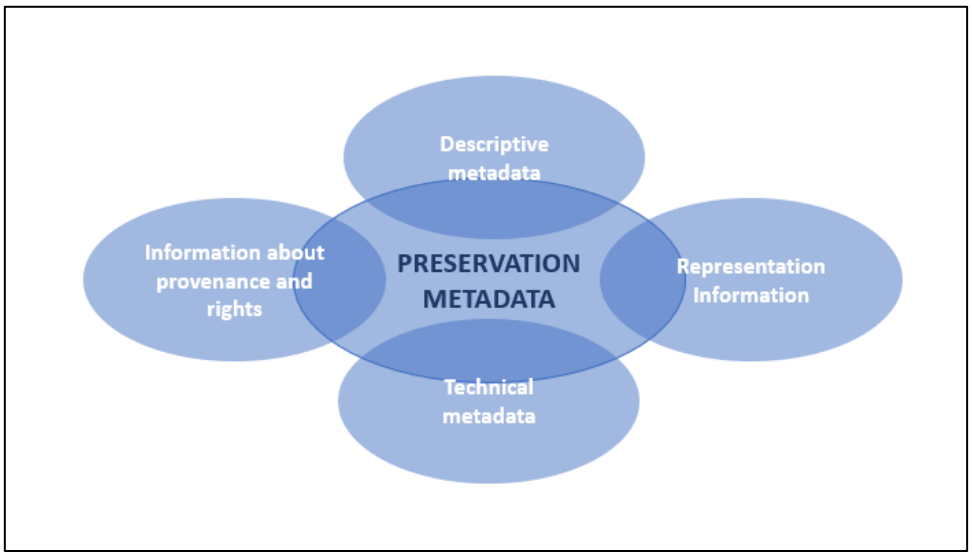

Figure 1: Preservation metadata as subset of all the information that is needed for preservation.

[1] 'Preservation Plan 2019-2022', https://www.kb.nl/sites/default/files/docs/preservation_plan_2019-2022.pdf
[2] Consultative Committee for Space Data Systems, "Reference Model For An Open Archival Information System (OAIS)", June 2012, https://public.ccsds.org/pubs/650x0m2.pdf
[3] Figure 1 is the KB's view on preservation metadata, based on the visualisation of P. Caplan, 'Understanding PREMIS' (2009), http://www.loc.gov/standards/premis/understanding-premis-rev2017.pdf (page 7).

According to PREMIS, a digital object can be described on four levels: Intellectual Entity, Representation, File and bitstream, as shown in figure 2[4]. In order to render an Intellectual Entity, all the files – together forming at least one version of the Intellectual Entity – must be identified, stored and maintained. This is essential for user display of an Intellectual Entity. The representation is the set of file objects needed to render an Intellectual Entity. It is important to know the composition of a digital object so that the digital object is displayed correctly and is also shown in the right order if applicable. The components of a representation are therefore documented as structural metadata.[5] For example, an Intellectual Entity can be a book that consists of three representations: A preservation Master copy in high quality (TIFF), a modified master copy in a lesser quality (JP2) and an Access copy (PDF) that is meant for direct user access. A representation may consist of more than one file if this is necessary for correct rendering of the object, for example in the case of pages stored as separate files. If the Intellectual Entity happens to be a movie, we could conceivably have a representation with video, audio and subtitles as separate files. You will need all three of them to be able to display the complete representation of this Intellectual Entity. The representation, therefore, could consist of one or more files, that in turn can contain one or more bitstreams. A bitstream is a datatype within a file, that serves a specific rendering purpose, in a way just like the representation itself. The difference is that bitstreams are datatypes present within a single file instead of being a composition of multiple files.

Now that the different levels on the left side of figure 2 are explained, it is time to have a look at the other component of the PREMIS Data Model, shown on the right. Events contain information about actions taken on objects in the repository. They demonstrate the authenticity of the object. Agents have an acting role in events and right statements. Agents can be people, organisations or software applications. The specific role of the agent should be recorded. The rights section provides information about rights and permissions that apply to the objects in a repository. A preservation strategy may include, for example, creating copies of the material. The rights metadata describes which licences are granted to the repository so that the repository can take all the necessary actions to preserve the object in compliance with copyright law.

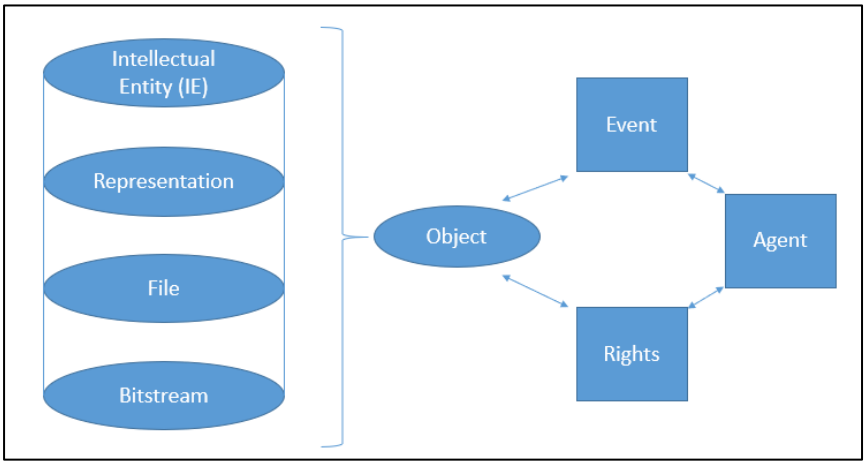

Figure 2: PREMIS Data Model.

[4] Figure 2 is based on Caplan, version 2 of PREMIS. See: 'Understanding PREMIS', page 6.
[5] Caplan, 'Understanding PREMIS'.

Preservation policy and preservation themes

To us, digital preservation means that we intend to keep digital information available while preserving the integrity and authenticity of the information. In short, this means guaranteeing that information is accessible, complete and authentic. Authentic in this sense means that data can be traced back to the digital material as it was originally received. Any changes to the data must be documented and performed in a controlled way in line with the preservation policy.

Preservation is not just a passive process of keeping data on reliable storage. It is not enough to store digital information and prevent it from change. After all, we want to be able to guarantee that material is safely stored, but also that is remains accessible. To facilitate reuse of the data, it might be necessary to change the format while preserving the content. This is the difference between bit preservation (where content and format never change) and functional preservation (where the content basically remains the same, but the format may change). Functional preservation is primarily about the content and not the form. Since it is not always possible to keep the content completely intact, it is important to determine significant properties that need to be preserved between transformations. Preservation of digital material is a living process. Long-term accessibility can be sustained because actions have been defined that counter information loss caused by technology changes and other contingencies. Three main themes running through our preservation policy are integrity, authenticity and long-term accessibility. These key concepts within our policy are heavily dependent on preservation metadata, so we will briefly summarize them here. It is important to realise that authenticity and integrity are a determining factor for long-term accessibility. The latter will be worked out in detail from page 5 onwards. We will also explore the relation between accessibility and preservation metadata.

Being able to prove that an object is what it appears to be, is an important part of authenticity. Based on file extension a file may appear to be in a certain format while the data does not comply with this format. This may occur when a file is inaccurately named. The actions that are performed on an object must be in line with established procedures that safeguard the preservation of intention, source and provenance. To put it differently we need to ask ourselves the following: can we verify that an object accords with the material that was supposed to be delivered? Do we know who delivered the object? Is the history of actions performed on the object stored as events within the metadata to provide users with a view on the digital lifecycle?

By integrity we mean guaranteeing that objects and collections are complete and that changes to the data are managed in a controlled and documented way. There are different aspects that determine integrity. Bit integrity means we can prove that a digital file is identical to the original publication. It is possible however, that multiple versions of a publication are present. For example, because an error has been fixed or because the producer uploaded a new version. Version integrity means making sure that the relation between versions is preserved and that these versions can be traced back to the original publication. Information Package integrity is important for making sure the package includes all the separate files that were part of the publication as it was delivered. The term information integrity is used for determining that all the representation information is preserved, either stored as separate documentation or as part of the metadata.

Finally we use the term collection integrity to ascertain that the collection is complete by verifying that all objects that should be present within a collection have a final availability status.

Preservation metadata and long-term accessibility

Above we briefly sketched the most important topics within the KB's preservation policy. An important part of digital preservation is being able to provide proof of things done in a trustworthy way. This evidence is stored in the metadata describing the digital objects. In this way an archive is able to show the data is safely preserved. Different aspects of long-term accessibility need to be covered in order to be considered trustworthy. Digital material must be findable, readable, interpretable, reliable and accessible in accordance with the guidelines for trustworthy digital archives (ISO-16363). Below we will detail these five aspects and their relation to preservation metadata.[6]

Findable

Digital objects can be considered 'findable' within the scope of a search interface. This requires descriptive metadata. Descriptive metadata contributes to the preservation theme of long-term accessibility.[7] Users must be able to find digital objects, for example by searching on a specific field like 'title', 'publication date' or 'author'. The term user is defined here in the broadest sense and means both external customers using KB services and internal users (employees) or even KB systems using system interfaces, data or metadata. All types of users need sufficient, reliable metadata[8] for querying and reporting. Internal users need this metadata to verify that digital objects are complete and may use it for reporting back to producers in case of errors or exceptions.

Persistent identifiers are also important for findability. These identifiers may be assigned by the KB (URN:NBN) or delivered as part of the metadata by producers (DOI). Because persistent identifiers are being maintained, a user can be sure that the identifier will always reference the same material, regardless of any changes to systems or user interfaces. Because of this, persistent identifiers contribute significantly to the findability of digital objects for the long term.

Readable

By 'readable' we mean being able to render and inspect digital objects. This aspect can only be guaranteed by building up knowledge on the types of file formats present in the archive. This knowledge can be used for risk analysis. For instance when considering our web collection, we may run the risk that older websites are not rendered correctly in modern browsers. Based on risk analysis a preservation strategy can be defined to mitigate this risk, for instance migration or emulation. In order to define a strategy like this we need technical information of all the different file formats, stored as metadata. This information forms a knowledge base consisting of three levels: stored (no file format identification), identified (basic file format identification) and known (extensive technical metadata). The 'known'-level is a requirement for considering migration to a new format because only then enough knowledge of the content is available to make informed decisions.

Technical metadata is therefore important for readability. All technical characteristics belonging to a file, like size and format can be considered technical metadata. When considering technical metadata, we must realise that building up complete and up-to-date knowledge is the main

[6] As described in the DUTO definition of long-term accessibility:
https://www.nationaalarchief.nl/archiveren/kennisbank/duurzaam-toegankelijk (version 19-07-2021).
[7] 'Preservation plan 2019-2022', page 7-9.
[8] To investigate how to define 'sufficient' we intend to implement different methods for analysing metadata quality. In this way we can work towards a process of gradually improving metadata based on user requirements and preservation goals.

function of this type of metadata, as opposed to provenance metadata where being able to reconstruct the history of an object is the most important function.

Interpretable

By 'interpretable' we mean that it is clear to users what it is they are looking at. Users should be able to understand the material, assisted by context information. Preservation metadata may have a specific function, for instance when certain metadata fields are used to support system functions. This is why we need to explain the purpose of available metadata fields: How are fields being used, and what goal do the fields serve? Preservation metadata and representation information help users in determining how objects must be interpreted. Structural and descriptive metadata can capture the relation between different objects, different versions of objects and the files within objects. This is important for determining Version integrity and Information Package integrity.

The aspect 'interpretable' is also important for functional preservation since understanding digital material is a key component here. To support functional preservation, we need to know what material we are dealing with, how it is classified and what a user can do with it. This information is stored as Representation information, that consists of context information linked to the material to make it more understandable, in the sense of technical composition as well as subject matter. Understandability is also improved by storing metametadata. This is information about the metadata itself, such as the origin of the original metadata and the standards used.

Reliable

If material is 'reliable', we mean that users can verify the digital material is complete and authentic. The life cycle of a digital object is documented in the preservation metadata, as provenance metadata. This part of the metadata describes how the material was delivered, what happened during or after ingest and whether any changes took place. It is important to understand that this life cycle starts even before ingest into the permanent repository. From the very first moment a digital object has been delivered to the storage, we need to record what happened to it. This history can be reconstructed from events, which contribute to the authenticity[9] of the object. The authenticity of a digital object is also supported by source metadata. This part of the preservation metadata contains, for example, the original metadata that was provided to us by the publisher along with a digital publication. The source metadata helps to determine whether the digital material really is what it seems to be. Files describing the delivery are also stored for verification of compliance with agreements between the archive and the publishers. In the ingest process, we also make sure only authorized persons have access. In the metadata a capture event is created for the delivery.

A completeness check on the digital material serves the preservation goal of integrity.[10] We speak of completeness on three levels: file, package and collection. The digital material that is delivered to the archives, is a Submission Information Package (SIP). The package that is stored in the permanent repository is an Archival Information Package (AIP). If a user consults the material from the permanent repository, we deliver a Dissemination Information Package (DIP). During these different stages, an information package will be checked for completeness. These checks are also available as an event in the metadata and tell us someting about IP-integrity: Are the expected files indeed present within the

[9] 'Preservation plan 2019-2022', page 6-7.
[10] Ibidem, p. 4-5.

SIP? And can the files within an AIP or DIP be traced back to the original SIP as it was delivered? The composition of the Information Packages in the different stages is documented in IP-schemas.

The checksum in the metadata also serves an aspect of integrity on file level, namely bit-integrity. The checksum can be used as input for a bit-integrity check, to make sure that the current copy of the material is bit-for-bit identical to the file stored in the archive. Checks on the AIP are done to confirm that the files did not become corrupt while stored in the permanent repository. Users have access to these checksums. In the context of accessibility, it is important that users can use the checksum to verify the bit-integrity of the data.

Available

Digital material can be considered 'available' when it is possible for users to receive a copy of the material. This also includes metadata. Availability of the metadata is crucial for internal users when making a risk analysis. However, for external users it is also important to have access to preservation metadata. As described above, preservation metadata can aid the user in determining if an object is authentic and complete, and making sure that any changes are verified and documented. All the information required for understanding an object (not only in a technical sense, but also for clarification of the content itself) needs to be available to users. Availability and accessibility are often used as terms that signify users always have open access to material, but copyright must be considered as well and therefore this is not always possible. But even if not all users have the legal right to access it, material can still be technically available. Furthermore, internal users always need access to the material for maintenance and administration. For the other aspects of long-term availability copyright also may apply, meaning that some users will not be entitled to use the material. This means that documenting rights is important, and rights information should be included in the metadata in full. Information about the copyright holder, licenses and conditions for reuse should be available to users. Metadata should even be available when digital material has been deleted. In that case a 'tombstone record' would be provided to users. This is a basic metadata record, informing the user at least about the deletion of the content.

Conclusion

Preservation metadata is an important part of the three preservation themes as described in KB's Preservation policy. Authenticity and integrity are preconditions for long-term accessibility. To guarantee access to digital material for the long term, the material must be retrievable, readable, interpretable, reliable, and available. When this is not the case, it is crucial that processes are in place for improvement. For instance, to guarantee readability a preservation strategy may be implemented. And in order to keep material available, functionality for downloading data and metadata could be developed. For the other aspects, it is mostly a matter of creating additional metadata to improve long-term accessibility.

Not only does preservation metadata help in guaranteeing digital material will be accessible for the long-term, the metadata also functions as evidence that the stored material in an archive is trustworthy. The life cycle of digital objects can be retrieved from the metadata, making it possible to track down potential changes. In this way objects can be traced back to their original form.

When the purpose of preservation metadata is documented, it can aid decision making regarding metadata. During the forthcoming migration of our digital collection for example, the KB will have to make choices on which metadata fields need to be migrated and why. When you understand the function of different types of metadata, it can be easier to make these choices.

Finally, the standards for metadata are important guidelines, but one must always keep the goal of preservation metadata in mind within the context of the institution. As depicted in figure 1, we have our own interpretation of preservation metadata based on PREMIS and OAIS, tailored to the context of the KB. In this manner it fits perfectly within the framework of the KB preservation policy. The KB is an institution without legal deposit, but with a strong focus on the users of the archives. To prevent that specialist knowledge about preservation remains locked within our own institute, we value connecting with preservation networks. In this way guidelines, goals and functions can be compared and transfer of knowledge between memory institutions is achieved.