Chapter 1
An Introduction to Implementing Digital Preservation Metadata

Angela Dappert, Sébastien Peyrard and Rebecca Squire Guenther

1.1 Introduction

Digital Preservation Metadata for Practitioners: *Implementing PREMIS* is written for anybody who needs to care for digital assets in any form over a long period of time. There are many decisions involved in this task, such as choosing data storage, backups and replication for safekeeping, or choosing file formats that can be read in the future. One key challenge that is addressed in this book is the question of what information one needs to keep, together with one's digital assets, so that they can be understood and used in the long-term. In other words, what metadata does one need?

Metadata are structured data that describe information objects, such as books, images, and maps, but also other objects. They are a key vehicle for accessing, managing, and understanding these objects. The properties that they describe are carefully chosen so that they are most helpful for the tasks they need to support. Librarians use metadata to create catalogs that help readers discover collection items by a variety of search criteria. The title of a book is the prototypical piece of metadata that comes to most people's minds. It helps you find the book and it helps you to decide whether it might be of interest to you. Online shoppers use metadata

A. Dappert (✉)
The British Library, 96 Euston Rd, London NW1 2DB, UK
e-mail: angela.dappert@bl.uk

S. Peyrard
Bibliothèque nationale de France, Site François-Mitterrand,
Quai François-Mauriac, 75706 Paris Cedex 13, France
e-mail: sebastien.peyrard@bnf.fr

R.S. Guenther
Consultant (Formerly Library of Congress), 215 W. 75th St.,
New York, NY 10023, USA
e-mail: rguenther52@gmail.com

© Springer International Publishing Switzerland 2016

A. Dappert et al. (eds.), *Digital Preservation Metadata for Practitioners*,
DOI 10.1007/978-3-319-43763-7_1

to find items to purchase and to decide whether the item meets their needs. Store managers use metadata to control their stock and to identify which items are in high demand. Photographers use metadata, embedded in digital images, to trace the history of an image or to inform editing decisions based on technical image information. In this book, we focus on metadata for a broad range of information objects.

1.2 Digital Preservation Metadata: Useful Information for Long-Term Access to Digital Objects

In addition to these search, discovery, access, rights, management, provenance, or technical metadata, we need to ask ourselves what metadata we need in order to keep digital information objects accessible over a long time—that is, to ensure their digital preservation. One mostly does not need to think about how to open a digital image, access an internet page, or edit a document when one accesses it on the same generation of computer that was used to create it. The file and the computer environments are compatible, the software needed to render or execute the file is installed, licenses are current, the software is supported by the software vendor, and you do not need much additional information about the file or the software or hardware it depends on. This is not the case if the data carrier on which the file is stored is unusual or outdated; if the file format is proprietary, rare or older; if the software or hardware is no longer supported; or if the digital object has undergone changes over time.

If we want to ensure the long-term usability of digital objects, it is necessary to gather enough information so that we can keep them accessible in some form in the future. This information is referred to as digital preservation metadata. This is particularly important for *repositories*, places where information objects are stored and managed for a long time. Simply storing digital objects on a data carrier is not enough to keep them usable. They need to be managed in a repository so that they are protected from accidental or intentional damage and so that a full computing environment can be created in which they can be accessed and understood when they are needed.

1.3 Standards for Digital Preservation Metadata

Over the last decade independent community activities emerged that initially defined their own metadata needs. But it quickly became clear that it was much better if the main actors shared the effort and worked toward developing a shared standard. Experts from key memory institutions and repository developers joined together to form the PREMIS Working Group to do exactly this. *The PREMIS Data*

Dictionary for Preservation Metadata [1], a de facto standard, now defines *core* metadata that are needed by most preservation repositories. There is a large variety of information object content types, such as documents, images, audiovisual material, web pages, spreadsheets, and business management files in proprietary formats. One needs additional application or content type specific metadata that go beyond this core metadata, to achieve long-term usability of their specific features. These forms of metadata are defined in additional standards that can be combined with the PREMIS core preservation metadata standard.

Use of standards is important as it supports the development of a community of best practice; it helps you learn from the insights of others, so that you do not inadvertently overlook key metadata in your own practice; it allows for development of tools to make metadata creation and management easier; and it enables organizations to more easily exchange information with each other.

1.4 How to Develop a Digital Preservation Metadata Profile

Because standards are broadly applicable and flexible they need to be customized to fit the context of an individual organizational situation. The PREMIS Data Dictionary, as the de facto digital preservation metadata standard, provides a data model consisting of basic entities (objects, agents, events, and rights) and basic properties (called '*semantic units*') that describe them. Understanding the dictionary alone does not teach you how to create your preservation metadata. It is like a language definition. It gives you a basic grammar and vocabulary. But it does not give you the sentences that tell your story. You use its constructs in order to write your own story. You have considerable freedom in how you do this. You will not use all of the words in the language; you choose how to structure the world you are describing; you may define some custom vocabulary for your own specific domain; you may fall back on foreign languages to express specific parts of your story, you may create your own dialect or accent to make this language serve your needs.

The result of customizing a set of standards to your needs is called a *metadata* (or *application*) *profile*. An application profile can be defined as follows:

> A set of metadata elements, policies, and guidelines defined for a particular application. The elements may be from one or more element sets, thus allowing a given application to meet its functional requirements by using metadata from several element sets including locally defined sets [2].

This book helps readers understand which options need to be considered in specifying a digital preservation metadata profile so that it is customized to their individual content types, technical infrastructure, and organizational needs. It provides practical guidance examples and raises important considerations. It does not provide a full-fledged implementation solution that can, by definition, only be specific to a given preservation context. As such, the book forms the bridge

between the formal specifications provided in a standard, such as the *PREMIS Data Dictionary* [1], and a specific implementation. First, it explains the thought-processes needed to decide what digital preservation metadata are needed and how they should be organized. Once this step has been accomplished, we can turn to the task of identifying how the needed metadata can be obtained. Will it be extracted automatically from the digital objects or from the metadata provided by, say, their publisher? Will it be created manually? Will it be submitted in various forms from various sources and need to be brought into a uniform format? One can then go on to choosing the most suitable standards for implementing the metadata and then, finally, to determining how to implement the specification in the chosen standards and serialization format. All of these topics are addressed in this book.

The book gives an introduction to fundamental issues related to digital preservation metadata and then proceeds to develop an in-depth understanding of issues related to its practical use and implementation. It should be of use to beginners and current practitioners. It is equally targeted at digital preservation repository managers and metadata analysts who are responsible for digital preservation metadata, as it is at students in Library, Information and Archival Science degree programs, or related fields. It can be used at the conception stage of a digital preservation system or for self-audit of an existing system.

This book is usable independent of the chosen standard or the version of the chosen standard. Rather than giving instructions on how to implement with a specific version it is about how to specify and implement your own digital preservation metadata profile to match your content type and organization. At the point of publication of this book PREMIS version 3.0 has been released but most existing implementations still use earlier versions. This is a normal aspect of using a standard since standards develop with user needs. Examples in this book are given in a specific version but usually can easily be translated to newer versions.

1.5 Reading Guide to This Book

Chapter 2 explains how risk and requirements analysis methodologies can be used as the basis for determining the required metadata.

The PREMIS Data Dictionary is generally applicable to all digital preservation scenarios, but it cannot give specific solutions on how it should be implemented for a particular application or for a specific content type. This means that you need to create a customized, content type specific and organization specific profile that specifies the required entities, structural relationships between them, and their properties. In other words, you define a specific profile of the generic Data Dictionary. In fact, this profile specifies those components that are *required* for your specific scenario based on a risk and functionality-driven requirements analysis.

Preservation risks are, among others, organizational, policy, economic, legal, or technological risks. The risk mitigation strategies that will be implemented in the repository depend on the availability of the appropriate metadata.

Requirements can also be defined based on

- the basic functions to be performed on a repository,
- the need for integration with existing systems, and
- the need for integration with related existing metadata and their standards.

The requirements analysis should always be agnostic of the eventual implementation solution and focus just on the required functionality. But sometimes the metadata choice is determined by what types of workflows can be implemented, and whether they may have to be manual or automatic. There are also nonfunctional requirements concerning cost and efficiency that can affect the choice of implemented metadata.

Requirements analysis in the domain of digital preservation does not need to start from scratch. Frameworks exist that

- describe the current best practice of digital preservation functionality that is to be supported in a repository, such as OAIS [3];
- an understanding of the basic preservation goals that need to be achieved, such as the ones defined by Caplan [4];
- risk analysis approaches, such as SPOT [5] or DRAMBORA [6]; and
- agreed core digital preservation metadata, such as is defined in the PREMIS Data Dictionary [1].

They all can inform the requirements definition process, but all of them need to be customized to the specific situation.

Chapter 3 explains the principles behind the PREMIS Data Dictionary and puts it in the context of the OAIS model.

The Data Dictionary specifies

- the data model,
- the basic categories of preservation metadata,
- the principles behind its design,
- how to apply it in practice, and
- the bodies and activities needed to ensure that the standard evolves together with its user community's needs.

OAIS [3] is a fundamental framework for long-term repositories, but PREMIS goes beyond OAIS in supporting the whole life-cycle of digital objects.

Chapters 4 through 12 explain the methodology for designing an application profile and illustrate implementation choices that have been made by leading institutions from around the world for specific entities and content types. This discussion is technically neutral and illustrates a variety of aspects that need to be considered in their context.

Chapter 4 discusses the general methodology in designing the specific logical data model for the context. Once the metadata requirements are known, one can specify the data model so that it supports the implementation of these requirements. The PREMIS Data Dictionary defines the general entities of the basic data model.

One needs to determine how they are to be tailored to the specific scenario. Which entities are significant and implemented depends on the functionality that is to be supported through the metadata. For example, for an '*e-journal*' scenario there may be '*journal*', '*issue*', '*article*', and '*figure*' objects that are particular subentities of the PREMIS Object. Do they all need to be implemented? For which of them do we need to create *Intellectual Entity* objects, so that we can capture descriptive metadata for them, to support search and access? Which of them have concrete digital realizations in the form of *File* or *Bitstream* objects that need to be described by technical metadata? Which of those have *Representations* that consist of several files for a single rendition of the object? Which events, agents and rights need to be captured to provide evidence of the digital assets' authenticity over time? Additionally, one needs to determine how the chosen entities are related to each other. Having access to an '*article*' object, for examples, makes it possible to follow its linking relationships to related '*prepublication*' objects and to the events that have affected this object over time.

Rather than defining a data model from scratch one can also reuse other people's profiles or customize default profiles that come with a digital preservation software solution. The general methodology for creating an application profile can equally be applied to this task of customization.

The case studies in Chaps. 5 through 12 illustrate how '*object*'-specific issues have been decided

- for different *entity types*, such as objects, events and rights;
- for different *content types*, such as web archives, audiovisual or e-book materials; and
- for different *organization types*, such as archives as opposed to libraries.

They include the choice of data models; the needs of the designated user communities; purposes in different communities, such as archives, libraries, museums; purposes of the collections; the functions the metadata need to support, such as storage, search, browsing, access, exchange, data management or preservation actions; intended scales for the size of the collection, IT resources, and human resources; what is particular about the content type that might impact the way you implement PREMIS descriptions of it; what other metadata systems were integrated and how that influenced PREMIS choices; what is in the scope of PREMIS and what is not; how the repository architecture influences metadata decisions; what policies or regulations affect implementation choices; pros and cons of choices; pitfalls and lessons learned. Chapters 11 and 12 discuss the use of event, agent and rights metadata.

Chapter 13 explains the serialization options, with XML, RDF and relational database implementations as examples of three common choices.

Once the basic entities in the logical data model and their relationships are defined, and the relevant properties that describe them are decided, one can go on to design the physical data model. A serialization is

> the process of translating data structures or object state into a format that can be stored (for example, in a file or memory buffer, or transmitted across a network connection link) and reconstructed later in the same or another computer environment [7].

The serialization of the metadata depends on the chosen metadata standards. On the other hand, a preferred serialization choice may influence which standard to choose.

It is important to note that PREMIS is completely implementation independent. It is a data dictionary—that is, a way of organizing your domain model with applicable semantic units that describe the entities in the model. It helps you think about your domain and its requirements. It does not at all specify how you implement it and almost everything is optional rather than mandatory—so that you can choose only what is needed by you.

So how, then, do we decide what serialization to choose? How do we combine different standards? In what way are different metadata uses supported by different serializations? And how can one reuse existing schemas and controlled vocabularies? Different choices of serialization have different advantages and disadvantages and support different functionalities. Factors that are relevant for the decision-making are: how serialization choices for multiple implemented standards complement each other; available IT skills; scalability problems; response times; how indexed metadata and administrative metadata that are associated with individual digital objects support search of the content; and the impact of the ubiquity and popularity of the serialization type on its tool support and continued usability. Optimization potential can be important, for example, to avoid repeating shared information over and over. In addition, existing and planned storage solutions are closely linked to serialization choices and impact data management and the way metadata are created, read, updated and deleted.

Chapter 14 discusses how to make different metadata frameworks work together in an institutional ecosystem.

It was already mentioned that it is generally necessary to combine several metadata specifications in order to support the complete functionality of the system. Extension schemas let you embed content type or file format specific metadata within a more general, core metadata framework. Metadata container formats, such as METS [8], have the specific purpose of tying different metadata frame works together. And packaging container formats are intended to package multiple content objects together with their metadata into an aggregate archival file, such as WARC [9]. Domain-specific metadata schemas that are typically descriptive and are used in a specific context, such as for libraries or archives, can be used to complement core preservation metadata.

When combining different frameworks one has to consider how to deal with possible redundancies when the same information is repeated in several places, and how this may affect the scalability, processing, versioning, and synchronization of the redundant information.

Chapters 15 through 17 discuss tools and systems.

Digital asset repositories can be developed in-house from scratch; or they can be commercial or open-source systems that are customized to varying degrees; or they can be commercial services, possibly offered in the cloud. The choices are determined by a variety of factors: What degree of customizability is required? What amounts of digital objects need to be accommodated? How does the solution fit in with the existing infrastructure? How much in-house IT support is available? What are the specific characteristics of the digital objects, in terms of size, access, or viewer requirements? What cost models are most appropriate for the intended use patterns? All of these requirements may determine the choice of implementation solution. Typically the system architecture is modular and layered and several system solutions are combined to accommodate all the required functionality. This includes

- pre-ingest functions, such as virus checking and metadata creation, through metadata extracting characterization tools;
- preservation actions, such as the creation of derivative formats;
- workflow issues;
- authorization and authentication;
- asset management, including the management of persistent, unique identifiers, the creation, update or deletion of digital objects or their metadata, their versioning, and the assurance of their integrity;
- persistent storage in nonproprietary forms;
- the creation of indexes to support discovery or search;
- cross-walking between different metadata formats;
- access through browsing, search, facets, or appropriate visualizations; and finally the
- viewing or delivery of the digital object.

For all of these life-cycle functions various tools may be combined into one system. They may be supported by registries that share important information, such as

- file format registries that help identify the file types of the files in the repository;
- controlled vocabulary registries that share vocabulary for enumerative data types in the domain, such as checksum algorithms in use, or typical preservation event types;
- software registries that describe, or even preserve, the software that is needed to render or execute a file.

Many of these tools, systems, and registries implement PREMIS and are shared by the digital preservation community.

Chapter 18 discusses what it means to be conformant with PREMIS.

It was said above that, whenever possible, repository systems should produce metadata that conform with standards. Conformance and compliance are however not always straightforward, especially if there are requirements to keep a standard as flexible as possible, so that it is widely usable.

1.6 Conclusion

Practitioners are sometimes intimidated by the prospect of having to specify their metadata profile. In the end metadata exist in order to support the implementation of system functions. Their specification is a part of the overall system requirements and implementation process. Tools and the metadata that support them need to meet your needs. Even if you cannot foresee the future and cannot now know what metadata you will wish you would have kept, it is better to get started. With the help of established systems and standards you can create satisfactory solutions from which you can learn and continuously improve your organization's services. Additional support comes from community development, through user groups, mailing lists, implementation registers, and reusable profiles. There is a lively user community of people who are concerned about the long-term preservation of digital objects. Practitioners are not alone.

References

1. PREMIS Editorial Committee (2015) PREMIS data dictionary for preservation metadata, version 3.0. http://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf. Accessed 24 Apr 2016
2. Dublin Core Metadata Initiative (DCMI) (2001) Dublin Core glossary. http://dublincore.org/documents/2001/04/12/usageguide/glossary.shtml#A. Accessed 24 Apr 2016
3. Consultative Committee for Space Data Systems (2012) Reference model for an Open Archival Information System (OAIS). CCSDS Secretariat CCSDS 650.0-M-2. Magenta Book, Washington, DC, Issue 2, June 2012
4. Caplan P (2008) The preservation of digital materials. Libr Technol Rep 44(2):9
5. Vermaaten A, Lavoie B, Caplan P (2012) Identifying threats to successful digital preservation: the SPOT model for risk assessment, D-Lib Magazine, September/October 2012. http://www.dlib.org/dlib/september12/vermaaten/09vermaaten.html. Accessed 24 Apr 2016
6. DCC (2015) DRAMBORA: Digital Repository Audit Method Based On Risk Assessment. Welcome to DRAMBORA interactive: log in or register to use the toolkit. http://www.repositoryaudit.eu/. Accessed 24 Apr 2016
7. Wikipedia (2015) Serialization. https://en.wikipedia.org/w/index.php?title=Serialization&oldid=715389105. Accessed 24 Apr 2016
8. Digital Library Federation (2010) <METS> Metadata Encoding and Transmission Standard: primer and reference manual, version 1.6 revised. http://www.loc.gov/standards/mets/METSPrimerRevised.pdf. Accessed 24 Apr 2016
9. ISO 28500:2009 (2016) Information and documentation—WARC file format, 2009. http://www.iso.org/iso/catalogue_detail.htm?csnumber=44717. Accessed 24 Apr 2016