Metadata Creation and Management

IMPACT Best Practice Guide

IMPACT project
Michael Day, British Library

Table of Contents

A Best Practice Guide to Metadata Creation and Management

Digitization does not equal access. The mere act of creating digital copies of collection materials does not make those materials findable, understandable, or utilizable to our ever-expanding audience of online users. But digitization combined with the creation of carefully crafted metadata can significantly enhance end-user access; and our users are the primary reason that we create digital resources (Baca, 2008, p. vi)

What is metadata?

General definitions of metadata frequently refer back to its literal meaning of "data about data", but most emphasise its role as a convenient way of referring to all of the many types of information that are needed to enable the retrieval, use and management of all types of information objects and collections. For example,

the US National Information Standards Organization (NISO, 2004) has defined metadata simply as any "structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage" any other resource. The extremely broad scope of these definitions means that theoretical discussions of metadata can sometimes seem daunting, but the main practical focus will always need to be focused on defining metadata in relation to its perceived role or function, and in particular to the specific use cases that it is intended to support (e.g. Day, 2005, p. 8).

There are various ways of categorising metadata. One popular approach first developed in the 1990s by the Making of America II Testbed Project (Hurley, *et al*. 1999) defined categories for descriptive, structural and administrative metadata types.

In this simple typology, *descriptive metadata* is that used for the discovery and identification of objects, *structural metadata* supports the discovery and navigation of objects, and *administrative metadata* includes any management information needed for the object, including information on the creation process, storage formats, the source and provenance of objects, and the intellectual property rights held in them (Day, 2005, p. 8).

While this typology has proved influential, not least on the Metadata Encoding and Transmission Standard (METS), it seems to be missing a clear understanding of the important role of context. Gilliland (2008) has preferred to categorise metadata with reference to its need to record both *intrinsic* and *extrinsic* properties of objects. She comments that metadata needs to reflect the *content*, *context* and *structure* of information objects.

Categorization of information object features from Gilliland (2008):

- Content relates to what the object contains or is about and is intrinsic to an information object

- Context indicates the who, what, why, where, and how aspects associated with the object's creation and is extrinsic to an information object

- Structure relates to the formal set of associations within or among individual information objects and can be intrinsic or extrinsic or both

The many functions that metadata can support means that there are a bewildering array of overlapping and interlinked standards and schemas. An approximate idea of their complexity can be gained by a quick look at the University of Montreal's MetaMap diagram, where information school students have provided a visualisation of international metadata standards and initiatives (as of 2005) in the form of a hypothetical subway-map. [1] However, only a relatively small number of these initiatives and standards will be relevant to the large-scale digitisation of text. The next section will introduce some of these standards in more detail.

General principles

Metadata can play many roles within large-scale text digitisation projects and programmes, covering the whole workflow from selection to the packaging of content for access (Table 1). The exact range and nature of the metadata required by projects will vary, as specific project requirements will reflect different aims and objectives and technical approaches. However, there are a number of general principles that might help projects develop suitable metadata strategies.

Metadata type	Potential role in digitisation programmes	Candidate standards

---

[1] MetaMap. Retrieved 19 January 2010 from: http://mapageweb.umontreal.ca/turner/meta/english/ [requires Adobe SVG viewer plugin]

Descriptive metadata at collection or item level	The selection of content for digitisation	Various (includes: MARC, MODS, Dublin Core, EAD, TEI Header, textMD)
	Supporting discovery and retrieval of the digitised content	
Identifiers	The *consistent* identification of content throughout the whole digitisation process	Various
	The packaging of content	
	Supporting access and reuse of the digitised content	
Technical metadata about images	Recording information about the results of imaging processes, e.g. file formats, colour spaces, compression algorithms, etc.	NISO Z39.87, MIX
	Recording information on image enhancements undertaken prior to OCR, image binarization, etc.	
Page layout metadata	Recording the text produced by OCR for a particular page, together with word and paragraph, text block and illustration co-ordinates	ALTO (a METS extension schema)
Text encoding	The identification of the structural elements of texts	TEI, TEI-Lite
Content packaging	Enabling complex packages of digitised content (e.g., multiple page images with associated OCR text, metadata) to be kept together for management and end-user access	METS, MPEG-21 DIDL, OAI-ORE
Preservation metadata	Technical information that can help support the longer-term sustainability of digitised content	PREMIS Data Dictionary, NLNZ Preservation Metadata, LMER (DNB)
Administrative metadata	Recording information on the digitisation process itself, e.g. documenting choices made on content selection, the use of specific tools for image enhancement or OCR, language dictionaries used, etc.	

Types of metadata and their potential roles in digitisation programmes

Understanding project requirements

The first general principle underlying metadata choices in any digitisation project is to understand specific project requirements. For example, projects are likely to have specific requirements relating to collection

usage and management, or resource constraints. Many of the choice factors identified by the JISC Digital Media guidance document on metadata (see the box below) relate to a deep understanding of project requirements.

JISC Digital Media guidance on factors likely to influence metadata choices (2009):

- **Your users and their needs** - what kind of information do they require and expect?

- **Your own needs as a collection manager** - what information do you require to manage, deliver and preserve your collection?

- **Your community's approach to metadata** - are there clear standards being used by similar collections?

- **Your legacy metadata** - what metadata already exists, what form does it take?

- **Existing systems** - does the metadata need to work well within particular systems (e.g. library catalogues, VLEs)?

- **Your resources** - how much time can you allocate to cataloguing; can you really afford to fill in dozens of categories or do you need something simpler?

- **The level of technical expertise available** - e.g. have you got staff who can understand XML?

- **Interoperability** - how important is it that your collection works alongside other collections?

- **The future development of your collection** - e.g. do you expect it to grow to include other formats or subjects?

From: JISC Digital Media, *Metadata* (March 2009). Retrieved 2011, from: http://www.jiscdigitalmedia.ac.uk/crossmedia/advice/metadata-standards-and-interoperability/

It is important that end users and their perceived needs are carefully considered at the project planning stage, as requirements identified here are likely to influence important decisions on delivery mechanisms and packaging formats, as well as on metadata. Another set of project requirements will relate to collection management needs, taking into account plans for future development and the need to consider the longer-term sustainability of digitised content. Other specific requirements will relate to the project context, e.g. with regard to the need for the project to interact with existing information systems, with legacy metadata, with third party systems and services.

Use existing standards, wherever possible

Another general principle relating to metadata is to use or adopt existing standards wherever possible. While it would be possible, in theory, for digitisation projects to design customised solutions to fit their specific metadata requirements, most choose not to do so. Adopting existing standards has a number of advantages. Firstly, it is likely to be much cheaper than developing your own metadata schema. Secondly, while specific project requirements are likely to differ in detail, there is likely to be some 'core' metadata functionality that will be common across multiple digitisation projects. Thirdly, using existing standards may provide additional opportunities for interoperability with other digitisation projects and may have a role in supporting the longer-term sustainability of content.

In practice, digitisation projects tend to use a wide range of different metadata standards, and rely heavily on being able to link relevant content and metadata together in content packages, typically using XML tools like the Metadata Encoding and Transmission Standard (METS).

Reuse legacy metadata, wherever possible

Another general principle is to capture and reuse whatever metadata might already exist, e.g. in library or publisher databases. For example, libraries will often have invested heavily in catalogues and databases that contain descriptive metadata about the physical items in their collections. Library catalogue data will usually contain a range of information types, typically standardised bibliographic descriptions supplemented by authority data (e.g., for personal and organisational names) and subject terms (e.g., subject headings or classification codes). In the more selective kinds of project, this information may often be useful in helping to decide whether items actually need to be digitised, e.g. by cross-checking with databases like the European Register Of Microform and digital Masters (EROMM).[2] Following digitisation, legacy metadata like this can also be packaged with content to support discovery and reuse, or could be supplied to third party services like EROMM or to rights registries, as required.

Automatically capture metadata, wherever possible

The industrial nature of large-scale text digitisation projects means that, wherever possible, metadata needs to be automatically captured as part of the digitisation process is something that needs to be taken account of in developing workflows. A key objective of projects will not just be to identify the specific types of metadata needed, but to identify how best this might be captured through the digitisation process. Examples of the kinds of metadata that can be captured in this way might include technical information about raster images (e.g., formats, colour depth, compression) or the linking of structural elements and OCR text with page image co-ordinates. Digitisation workflows will also need to take into account the automated generation of 'content packages' that are able to combine all of the elements that make up the finished product – including page images, OCR text, metadata of various types, maybe additional annotation layers – in a logical way.

Identifiers

Identifiers will be important at all stages of the digitisation process, and in particular to facilitate the integration of resource components - e.g., page image files, OCR generated text files - into the logical objects that will be needed by end users. In practice, identifiers will need to be applied at several different levels of granularity, e.g. for individual pages, for works and - in some cases - their component parts (e.g. book chapters, newspaper articles), and for aggregations of works at higher levels, e.g. to identify where content are part of particular collections or serials.

Persistent Identifiers

Several standard frameworks exist for supporting the unique and persistent identification of digital content. These are typically based on indirection, providing services that do not just assign unique identifier strings (e.g., on the ISBN model), but also include *managed* resolution, enabling identifiers to be matched with content over time. This means that persistent identification will incur ongoing costs (Hilse & Kothe, 2006).

All changes in location, ownership or metadata must be reflected in the [persistent identification] name-space system – causing the organisations that run an identification system to incur costs.

That said, it should be noted that persistent identifiers may not be needed at all levels within digitisation projects. They would be most useful where content (at any level of granularity) is intended to be shared across networks.

Further guidance on persistent identifiers can be found in a report published by the Centre for European Research Libraries and the European Commission on Preservation and Access (Hilse & Kothe, 2006) and in the Paradigm Workbook (Paradigm project, 2007).

---

[2]European Registry Of Microform and digital Masters; 1998-2011; http://www.eromm.org/ Retrieved 13.03.2011

The Digital Object Identifier System

Persistent identifier frameworks include the Digital Object Identifier (DOI) system, which provides an infrastructure for the persistent identification and resolution of all types of content object.[3] The DOI system is managed by the International DOI Foundation (IDF), an open membership consortium that includes both commercial and non-commercial partners, and is an implementation of the Handle System, part of the Corporation for National Research Initiatives's Digital Object Architecture.[4] DOI names are widely used in scientific publishing, where they are used to support resource discovery and citation linking through initiatives like CrossRef.[5] They are also beginning to be used to support the persistent identification of other kinds of content. For example the international DataCite initiative is exploring the role of the DOI System for the persistent identification of scientific data.[6] DOI names can be assigned at any level of granularity, e.g. in scientific publishing they can be used to provide identify at journal level as well as for individual issues and articles, and for particular components within articles, e.g. illustrations or tables. The IDF is also a member of the Thematic Partner Network of the Europeana v1.0 project, where it provides technical advice on identifiers and metadata.

Uniform Resource Names

Another framework for persistent identifiers is the Uniform Resource Name (URN), part of a wider architecture of Internet standards focused on the discovery, description and network location of digital objects. The URN syntax can be used to encode bibliographic identifiers like ISBN, ISSN and SICI, and has also been used as the basis of the National Bibliography Number (NBN) used by a number of national libraries, e.g. in the nordic countries, the Baltic states and Germany. Defined in Internet RFC 3188 (Hakala, 2001), NBNs are typically used to assign identifiers to objects that do not have publisher assigned numbers, e.g. for 'grey literature' or Web pages. The German National Library[7] and the National Library of Sweden[8] have both developed services that support the assignment and resolution of NBNs.

ARK Identifiers

A emerging, perhaps more lightweight, standard for persistent identification is the Archival Resource Key (ARK) maintained by the California Digital Library.[9] ARK does not focus on indirect naming, but is based on the idea that "persistence is purely a matter of service, and is neither inherent in an object nor conferred on it by a particular naming syntax" (Kunze, 2003). The ARK specification defines a special kind of actionable URL (Uniform Resource Locator) that links three things, the object, its metadata and the current provider's 'commitment statement.' The structure of the URL separates the unique object identifier itself (the Name Assigning Authority Number) from its actionable hostname (the Name Mapping Authority Hostport). If, therefore, the hostport fails to work or is no longer actionable through HTTP, the specification defines ways to realign the object with its Name Assigning Authority. ARK identifiers are used by many organisations including the California Digital Library, the National Library of Medicine, Portico and the Bibliothèque national de France (e.g., in Gallica).

Major categories of metadata

The main types of metadata needed to support large-scale digitisation are descriptive, metadata, structural and preservation metadata. This section will provide a more detailed introduction to all of these categories together with overviews of particularly relevant standards.

---

[3] Digital Object Identifier homepage; 2010; http://www.doi.org/ Retrieved 13.03.2011
[4] Handle System homepage; 2010; http://www.handle.net/ Retrieved 13.03.2011
[5] CrossRef homepage; 2010; http://www.crossref.org/ Retrieved 13.03.2011
[6] DataCite homepage; 2010: http://www.tib-hannover.de/fileadmin/datacite.html Retrieved 13.03.2011
[7] *Deutsche Nationalbibliothek, Persistent Identifier*. Deutsche Nationalbibliothek; http://www.persistent-identifier.de/ Retrieved 13.03.2011
[8] National Library of Sweden, *URN:NBN*; 2010: http://www.kb.se/english/about/projects/digital/urn-nbn/ Retrieved 13.03.2011
[9] California Digital Library, ARK (Archival Resource Key); 2009: http://www.cdlib.org/inside/diglib/ark/ Retrieved 13.03.2011

Descriptive metadata

The main role of descriptive metadata in large-scale digitisation projects is to support end user access and retrieval. Because most text digitisation initiatives run by libraries will already have invested deeply in metadata creation, the opportunity exists to capture much of the required information from legacy systems, e.g. library catalogues and databases, or publishers' databases. Publishers use an XML-based metadata standard known as ONIX to distribute information about books and serials within the book trade and with aggregators and libraries.[10]

Descriptive metadata in libraries tends to be a mixture of several different things. Library catalogues will typically contain bibliographic data derived from the works being described, supplemented by:

• Authority data, that helps to clarify standardised forms of dealing with things like author names (name authorities) or the linking of derivative items (like translations) to their parent work (uniform titles);

• Subject (or genre) data, typically applied from taxonomies like subject classification schemes or subject headings;

• Holdings data that will provide more detail on physical locations or local holdings (e.g. for runs of serials).

Library catalogues have tended to be non-hierarchical and the level of information held within library catalogues differs depending on the type of object being described. For historical reasons, books and monographs will typically be described at individual item level while serials (including journals and newspapers) will be described only at title level, with a record of holdings information. This means that there will be far more descriptive metadata available in library catalogues for digitisation projects focused on books than for those based on newspapers. In addition, it means that catalogues will often contain little or no metadata on specific book chapters or – even more significantly – individual articles in journals or newspapers. This means that digitisation projects may need to focus on developing automated means for users to be able to access digitised objects at lower levels of granularity than library catalogues have traditionally provided for.

Library cataloguing formats: MARC and ISBD

Many library catalogues use a set of descriptive standards that are structured around the MARC (Machine-Readable Cataloguing) formats first developed for bibliographic data representation and communication in the 1960s. In the Anglo-American world, the most widely used standard is the MARC21 format maintained by the Library of Congress. Elsewhere, IFLA's UNIMARC (Universal MARC) is widely used, sometimes as an internal catalogue format, but more often in its intended role as an exchange format supporting the international exchange of bibliographic records. Historically, MARC was based on its own transport format known as ISO 2709, but increasingly the maintenance agencies are providing tools that enable MARC records to be encoded in more 'Web-friendly' standards like the Extensible Markup Language (XML). For example, the Library of Congress Network Development and MARC Standards Office have developed the MARC 21 XML Schema as well as a number of conversion tools.[11]

The MARC formats merely define the broad structure of records and what are known as content designation (codes and conventions needed to manipulate records), while a group of separate standards determine the *content* of specific data elements. These standards include a set of generic rules for bibliographic records produced by IFLA, the International Standard Bibliographic Description (ISBD).[12] These are in turn embodied in influential standards like the Anglo-American Cataloguing Rules, Second Edition

[10]ONIX; 2010: http://www.editeur.org/8/ONIX/ Retrieved 13.03.2011

[11]MARCXML homepage; Library of Congress; 2011: http://www.loc.gov/standards/marcxml/ Retrieved 13.03.2011

[12]IFLA, ISBD Review Group; 2009: http://www.ifla.org/en/isbd-rg Retrieved 13.03.2011

(AACR2), which is soon to be superseded by the Resource Description and Access (RDA) standard.[13] The content of other MARC data elements - specifically those that identify the subject or genre of resources - will be based on a range of other standards like Library of Congress Subject Headings (LCSH) or the Dewey Decimal Classification (DDC).

Cataloguing formats: MAB2

Not all countries use the MARC formats. Since the early 1970s, Germany and Austria have used the MAB (*Maschinelle Austauschformat für Bibliotheken*) format as an exchange format for bibliographic records. The Deutsche Nationalbibliothek is the maintenance agency for the current version of this, MAB2-Format.[14] Like MARC, MAB is based in part on ISO 2709, but is more closely linked to the German cataloguing codes embodied in the *Regeln für die Alphabetische Katalogisierung* (RAK). An XML schema known as MABxml has been developed for the MAB2-Format.[15]

Since 2004, the Committee for Library Standards (*Standardisierungsausschuss*) has been working towards adopting MARC21 in place of MAB2, and its Data Formats Expert Group (*Expertengruppe Datenformate*) has explored in detail the differences between the two formats, working out what would need to be added to MARC21 in order to support existing data exchange activities.[16]

Interoperability: MODS and DCMI

A number of descriptive metadata standards have been specifically designed to support digital library operations. These will typically be lightweight schemes based on XML, primarily designed to support interoperability with other standards.

One example of these is the Metadata Object Description Schema (MODS), which - like MARC21 and other standards - is maintained by the Library of Congress Network Development and MARC Standards Office.[17] The standard can be used in many different ways: e.g. for producing original resource descriptions, or for representing simplified MARC records in XML format. It can also be used as an intermediate format for metadata conversion or union catalogues, and as a representation format suitable for metadata harvesting, e.g., using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH).[18] In addition, it can act as a means of producing metadata in XML that can be packaged with content, and is specifically designed to act as an extension schema to METS. MODS Version 3.3 has twenty top-level elements (mostly optional) and over sixty at other levels. Some commentators consider that this richness gives MODS significant advantages over other 'interoperability' schemas like unqualified Dublin Core (DC). For example, Gartner (2003, p. 8) has written that MODS "reconciles reasonably successfully the divergent demands of interoperability and precision which have caused problems for DC."

The Dublin Core Metadata Initiative (DCMI) is an international metadata standards organisation that has the aim of providing "simple standards to facilitate the finding, sharing and management of information."[19] The initiative is open to anyone that wishes to participate through domain-specific groups known as DCMI Communities. To date, DCMI's work on metadata standards has largely concentrated on two main things. The focus of the earliest Dublin Core workshops was the identification and development of 'core metadata' elements for generic resource description, resulting in a fifteen-element metadata element set that has been codified in a number of standards (including ISO 15863:2009) as well as the OAI-PMH. In parallel with this, the Dublin Core community has focused a great deal of attention on exploring ways in which

---

[13]Resource Description and Access homepage; 2011: http://www.rdaonline.org/ Retrieved 13.03.2011

[14]MAB2 homepage; Deutsche Nationalbibliothek; 2010: http://www.d-nb.de/standardisierung/formate/mab.htm Retrieved 13.03.2011

[15]MABxml homepage; Deutsche Nationalbibliothek; 2010; http://www.d-nb.de/standardisierung/formate/mabxml.htm Retrieved 13.03.2011

[16] Umstieg von MAB2 auf MARC21; Deutsche Nationalbibliothek; 2010: http://www.d-nb.de/standardisierung/formate/formatumstieg_herst.htm Retrieved 13.03.2011

[17]MODS overview: Uses and Features. Library of Congress; http://www.loc.gov/standards/mods/mods-overview.html Retrieved 13.03.2011

[18]Digital Library Federation, Aquifer initiative; 2010: http://www.diglib.org/aquifer/ Retrieved 13.03.2011

[19]DCMI Mission and Principles; Dublin Core Metadata Initiative: http://dublincore.org/about-us/ Retrieved 13.03.2011

metadata implementers would be able to extend DC using a wide range of namespaces and encoding schemes. Central to this is the concept of 'application profiles,' defined on the DCMI Web pages as "the idea that metadata records would use Dublin Core together with other specialized vocabularies to meet particular implementation requirements."[20] Influenced by the World Wide Web Consortium's Resource Description Framework (RDF), the DC community have focused recent effort on producing a metadata model (the DCMI Abstract Model) and framework that would help support the development of application profiles. The Singapore Framework for Dublin Core Application Profiles provides guidance on the design and documentation of implementation-specific profiles.

Dublin Core is used by a very large number of digital library projects, typically for dealing with interoperability challenges like cross-searching or metadata harvesting. It is sometimes used in digitisation contexts, again primarily as a means of integrating content from multiple sources. A prominent example is Europeana, which uses OAI-PMH to aggregate metadata about digitised resources from a vast number of partner organisations. Europeana does not harvest or store content itself. Instead it uses OAI-PMH to aggregate a defined subset of metadata from its partners, while directing users to the providers own sites for end user access. In order to support this, Europeana has developed a data model and metadata application profile that facilitates metadata integration between all project participants. Like other Dublin Core application profiles, the Europeana Semantic Elements profile combines existing DCMI terms with a number of additional elements deemed specific to Europeana's needs (Table 2).

Full integration currently depends on manually mapping providers' metadata fields to the Europeana Semantic Elements, which can involve a significant amount of effort. For example, in April 2009, the Europeana prototype contained metadata provided by over 50 institutions from 24 countries, including 15 different metadata formats (Concordia, 2009).

Technical metadata

Technical metadata is the term used to refer to any technical information about particular files or objects. Federal Agencies Digitization Guidelines Initiative's glossary provides some examples.[21]

For example, for digital photographs, this category includes the shutter speed and lens aperture; for all digital images, they include such things as pixel dimensions; for sound recordings, they include sampling frequency and bit depth. This subcategory is commonly embedded within image files using Exif, TIFF Header Tags and/or XMP, and within WAVE sound files using the AudioFormat chunk.

In the digitisation domain, technical metadata will relate to the image files produced by the image capture process, and in many cases could be captured automatically by the devices used or the digitisation workflow. In terms of standards, NISO defines a data dictionary for technical metadata for digital still images (ANSI/NISO Z39.87-2006), which has an XML representation defined in the NISO Metadata for Images in XML Schema (NISO MIX), a standard maintained by the Library of Congress' Network Development and MARC Standards Office.[22] The data dictionary itself defines a large number of elements, covering basic information about the file (e.g. file size, format designation, compression) and image (e.g. width/height, colour space and profiles), capture information (e.g. dates, capture processes), as well as format-specific information required for formats like JPEG 2000 or DjVu.

Technical metadata for text-based objects would normally be simpler than that needed for images, but could contain important information on things like languages or character sets (Gartner, 2008, p. 9). Relevant standards include the Text Encoding Initiative, where TEI Headers are be used to store technical metadata, and the smaller textMD element set and its matching XML Schema.[23]

---

[20]DCMI Metadata Basics. Dublin Core Metadata Initiative; 2010: http://dublincore.org/metadata-basics/ Retrieved 13.03.2011

[21]FADGI Glossary, "Metadata, Technical"; Federal Agencies Digitization Guidelines Initiative: http://www.digitizationguidelines.gov/term.php?term=metadatatechnical Retrieved 13.03.2011

[22]NISO Metadata for Images in XML Schema. Library of Congress; 2009: http://www.loc.gov/standards/mix/ Retrieved 13.03.2011

[23]textMD landing page; Library of Congress; 2009: http://www.loc.gov/standards/textMD/index.html Retrieved 13.03.2011

Structural metadata

The NISO introduction to *Understanding Metadata* (NISO, 2004) says that structural metadata "indicates how compound objects are put together, for example, how pages are ordered to form chapters." The Federal Agencies Digitization Guidelines Initiative's glossary adds that it "describes the intellectual or physical elements or a digital object."[24]

For a file that represents a single page as a compound document (e.g., a JPEG 2000 jpm file), the structural metadata may include information on page layout. In a multi-file digital object (e.g., a scanned book with many page images), structural metadata describes the object's components and their relationships: pages, chapters, tables of contents, index, etc.

The simplest example of structural metadata might be the table of contents of a book. However, where digitised texts are concerned, structural metadata is usually closely linked to packaging formats like the Metadata Encoding and Transmission Standard. These will be introduced in more detail in later sections of this guide. Yale University Library have produced specific guidance for the use of structural metadata by digitisation projects.[25]

Preservation metadata

Preservation metadata has been defined as any information that supports or documents the process of long-term digital preservation and specifically that information that "supports the viability, renderability, understandability, authenticity, and identify of digital materials in a preservation context."[26] In practice, it overlaps with most other categories of metadata, including structural, administrative and technical metadata. At present, the most mature metadata standard developed specifically to address long-term preservation requirements is the PREMIS Data Dictionary for Preservation Metadata, currently in version 2.0 (2008).[27] It is recognised that many digitisation projects may not primarily be concerned with long-term preservation, but the PREMIS Data Dictionary may provide some valuable pointers to those interested in the longer-term sustainability of content.

Source	Element	Refinement(s)
DC	title	Alternative
DC	creator	
DC	subject	
DC	description	tableOfContents
DC	publisher	
DC	contributor	
DC	date	created; issued
DC	type	
DC	format	extent; medium
DC	identifier	
DC	source	

---

[24]FADGI Glossary, "Metadata, Structural"; Library of Congress; 2009: http://www.digitizationguidelines.gov/term.php?term=metadatastructural Retrieved 13.03.2011

[25] Yale University Library, *Best practices for structural metadata*, v. 1 (June 2008); Library of Yale University; 2008: http://www.library.yale.edu/dpip/bestpractices/BestPracticesForStructuralMetadata.pdf Retrieved 13.03.2011

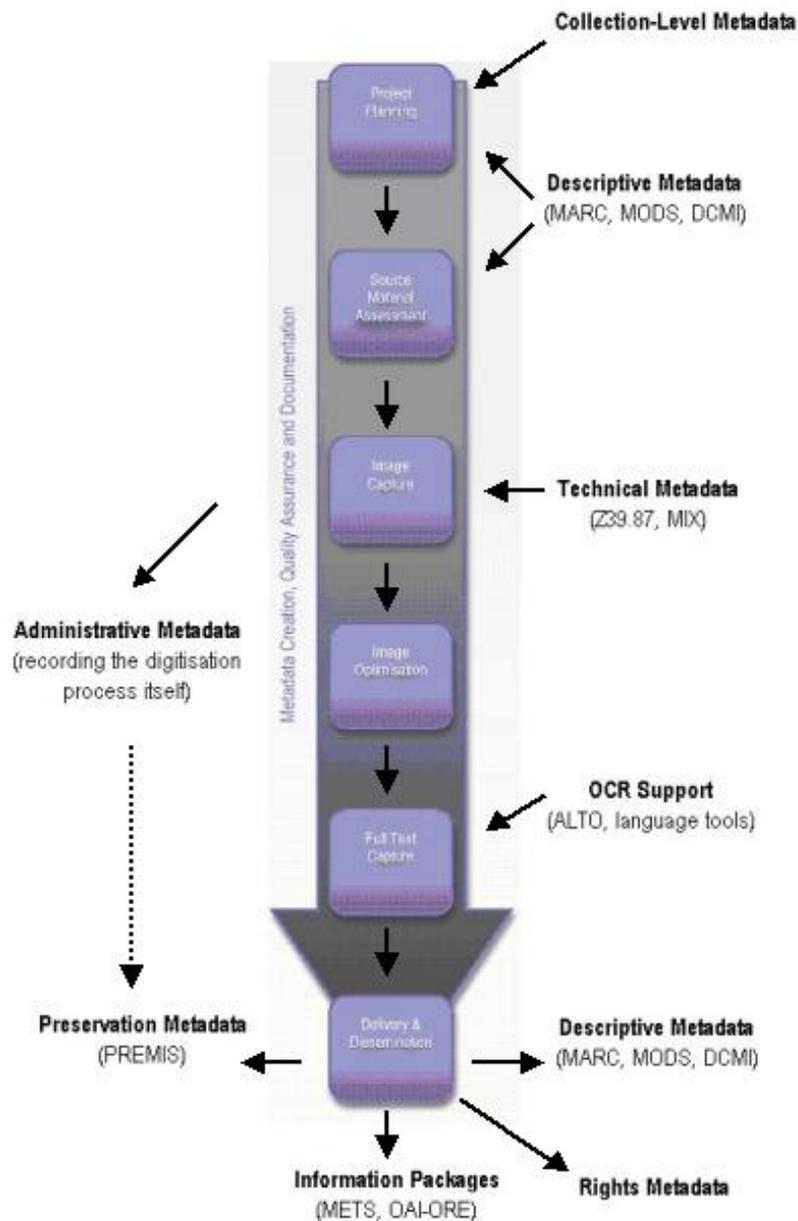[26]PREMIS Editorial Committee, *Introduction and Supporting Materials from PREMIS Data Dictionary for Preservation Metadata*, v. 2.0 (2008). Library of Congress; 2008-2011: http://www.loc.gov/standards/premis/v2/premis-report-2-0.pdf Retrieved 13.03.2011

[27]PREMIS Preservation Metadata; Library of Congress; 2011: http://www.loc.gov/standards/premis/ Retrieved 13.03.2011

DC	relation	isVersionOf; hasVersion; isReplacedBy; replaces; isRequiredBy; ...
DC	coverage	temporal; spatial
DC	rights	
DC terms	provenance	
Europeana	relation	isShownBy; isShownAt
Europeana	userTag	
Europeana	unstored	
Europeana	object	
Europeana	language	
Europeana	provider	
Europeana	type	
Europeana	uri	
Europeana	year	
Europeana	hasObject	
Europeana	country	

Europeana Semantic Elements. Source: Concordia (2009)

Metadata standards used in digitisation projects

The previous section attempted to introduce the wide range of different metadata types relevant to digitisation projects (Figure 1). This section will provide more detail on the metadata standards and frameworks that are actually used within the digitisation workflow.

Collection-Level Metadata

Descriptive Metadata
(MARC, MODS, DCMI)

Technical Metadata
(Z39.87, MIX)

OCR Support
(ALTO, language tools)

Administrative Metadata
(recording the digitisation
process itself)

Metadata Creation, Quality Assurance and Documentation

Project Planning

Source Material Assessment

Image Capture

Image Optimisation

Full Text Capture

Delivery & Dissemination

Preservation Metadata
(PREMIS)

Descriptive Metadata
(MARC, MODS, DCMI)

Information Packages
(METS, OAI-ORE)

Rights Metadata

Simplified mass digitisation workflow with major metadata types

Simplified mass digitisation workflow with major metadata types

Naturally, there are many options available for this, but the vast majority of text digitisation projects use one of two main standards: the Metadata Encoding & Transmission Standard (METS) and the Text Encoding Initiative (TEI) guidelines. The syntax of both of these standards is based on XML. The general approach of these two standards differs. METS is a generic means of packaging metadata, content and links together

to produce logical objects, and is widely used in a range of digital library contexts. In text digitisation projects, METS is typically used as a means of creating logical containers that are able to link all of the content files and metadata that make up a given work, can represent its structure (e.g. page order), and as a means of linking page images with OCR text, e.g. using extension schema like the Analyzed Layout and Text Object (ALTO) standard. The TEI guidelines, by contrast, were primarily designed for the detailed markup of texts, and its use in large-scale text digitisation contexts tends not to use all of its features.

The following sections will introduce both standards in more detail. Other potential packaging formats will be dealt with briefly later in this chapter.

The Metadata Encoding & Transmission Standard (METS)

The Metadata Encoding & Transmission Standard (METS) is a standard maintained by the Library of Congress's Network Development and MARC Standards Office.[28] METS is an attempt to provide an XML Schema for encoding metadata that can support the management and exchange of digital library objects. Essentially, it is an XML-based framework in which different types of metadata can be packaged together. Beedham, *et al*. (2005, p. 70) say that METS "uses XML to provide a vocabulary and syntax for identifying the components that together comprise a digital object, for specifying the location of these components, and for expressing their structural relationships."

A METS document currently consists of seven sections: a *METS Header* for brief descriptive information about the METS document itself, *Descriptive Metadata*, *Administrative Metadata*, a *File Section* listing all of the files that make up the object, *Structural Map* and *Structural Links* sections that enable individual files and metadata to be mapped to the structure of the object, and a *Behavior* section that provides information on how particular components should be rendered. The administrative metadata section is intended to store technical information about the file, as well as information about intellectual property rights held in the resource, the source material, and provenance metadata that records relationships between files and migrations. The modular design of METS means that objects can also include metadata from 'extension schemas' - i.e. from standards defined elsewhere. For example, the descriptive metadata could include or link to records conforming to standards like the Encoded Archival Description (EAD), the Metadata Object Description Schema (MODS), or Dublin Core. Technical information about images, for example, could be taken from the NISO *Data Dictionary - Technical Metadata for Digital Still Images* standard (ANSI/NISO Z39.87-2006), for example, in its XML encoding in MIX (NISO Metadata for Images in XML Schema).[29] Other extension schemas that could potentially be used by digitisation projects include ALTO (Analyzed Layout and Text Object),[30] and textMD (Technical Metadata for Text).[31]

METS in use

METS evolved from an XML Document Type Definition developed for the Making of America II digitisation project (Hurley, *et al*., 1999) and it has been most widely implemented to date in digitisation contexts (Gartner, 2002). It has been widely used, for example, in the Oxford Digital Library to provide integrated access to digitised image files with searchable texts.[32]

METS provides a general framework for the integration of various types of resources with their supporting information (metadata). Digitisation projects will need to decide which mixture of content and metadata standards are most appropriate. The potential complexity of these choices can be demonstrated by the National Digital Newspaper Program in the US (Littman, 2006).

---

[28]Library of Congress, Network Development and MARC Standards Office; Library of Congress; 2011: http://www.loc.gov/marc/ndmso.html Retrieved 13.03.2011
[29]NISO Metadata for Images in XML Schema (MIX); Library of Congress: http://www.loc.gov/standards/mix/ Retrieved 13.03.2011
[30]Analyzed Layout and Text Object (ALTO). Retrieved 19 January 2010 from: http://www.loc.gov/standards/alto/
[31]Technical Metadata for Text (textMD). Retrieved 19 January 2010 from: http://www.loc.gov/standards/textMD/
[32]Oxford Digital Library. Retrieved 19 January 2010 from: http://www.odl.ox.ac.uk/

An [newspaper] issue, including its sections and pages, are represented by a single METS record. Each section and page is described by a MODS record. For each page, there is a master image for preservation purposes (encoded as a TIFF), primary service image for online rendering (encoded as a JPEG2000), derivative image for downloading and offline use (encoded as a PDF), and OCR text for discovery (encoded with the Analyzed Text and Layout (ALTO) schema). Metadata for Images in XML (MIX) encoding of Technical Metadata for Digital Still Images (NISO Z39.87) and PREMIS metadata describe each of the images. Thus, each issue digital object is composed of a digital object encoding record (i.e., the METS record), which contains various metadata records (i.e., MODS, MIX, and PREMIS), and references various external digital object components (i.e., the TIFFs, JPEG2000s, PDFs, and ALTO files).

Choices on a similar range of standards and encodings – including TIFF, MODS, PREMIS, and ALTO have also been made (and documented) by the Australian Newspaper Digitisation Program (Lee, 2009).

Technical metadata for OCR: the ALTO standard

As has been noted, METS allows implementers to use XML extension schemas like textMD or the Analyzed Layout and Text Object (ALTO). ALTO is a XML-based standard that has been specifically designed to support digitisation projects that use OCR. Initially developed by the European METAe project and then maintained by CCS Content Conversion Specialists GmbH, the official maintenance agency for the ALTO XML Schema v. 2.0 is now the Library of Congress Network Development and MARC Standards Office.[33]

ALTO's main role is to provide technical metadata about the layout and content (e.g. OCR text output) of physical texts, typically recording the coordinates of images, text blocks, paragraphs and words on a given page. ALTO files have three main sections:

• Description - contains metadata about the file itself and information on how it was created

• Styles – includes information about text styles (fonts) and paragraph alignments

• Layout – contains the OCR content, subdivided into Pages (identifying the print space and margins)

When combined with METS, ALTO is able to provide a rich representation of the original object. For example, Digital Library Consulting says that this can have a major influence on users' search experiences.[34]

For example, a typical METS/ALTO object encodes not only the complete logical and physical structure of a document (i.e. chapters, sections, articles, pages, etc., and their associated metadata), but also the full-text content of each section of the document and even the physical coordinates of every word in the document!

METS/ALTO is used by a large (and growing) number of digitisation programmes, including many newspaper projects. A 2007 survey undertaken by the National Library of the Netherlands in connection with its Databank of Digital Daily Newspapers project showed that around half of respondents used ALTO for storing information about page zoning and segmentation (Klijn, 2008). METS/ALTO has the advantage that much of its key information can be automatically generated from the digitisation and OCR processes and workflow itself.

More information on METS/ALTO:

METS Official Web Site: http://www.loc.gov/standards/mets/ Retrieved 13.03.2011

This is the most authoritative source of information on METS, provided by the Network Development and MARC Standards Office of the Library of Congress, including links to tutorials and schema specifications.

---

[33] ALTO; Library of Congress; 2010: http://www.loc.gov/standards/alto/ Retrieved 13.03.2011
[34] "What's METS/ALTO and should you care?"; Digital Library Consulting; 2009: http://www.dlconsulting.com/blog/?p=46 Retrieved 13.03.2011

METS: An Overview and Tutorial: http://www.loc.gov/standards/mets/METSOverview.v2.html Retrieved 13.03.2011

JISC TechWatch Report. *METS: Metadata Encoding and Transmission Standard*. Gartner, R. (2002): http://www.jisc.ac.uk/whatwedo/services/techwatch/reports/horizonscanning/hs0205.aspx Retrieved 13.03.2011

ALTO Official Web Site; Library of Congress; 2010: http://www.loc.gov/standards/alto/ Retrieved 13.03.2011

The Text Encoding Initiative (TEI)

The TEI (Text Encoding Initiative) Guidelines are an encoding scheme that was originally designed for the detailed markup of scholarly texts. Since 2000, the guidelines have been hosted and maintained by an international membership organisation known as the TEI Consortium (Burnard, 2000).[35] First developed in the late 1980s with its first full release in 1994, the latest version of the guidelines is currently TEI P5 (Burnard & Bauman, 2007).

Introduction

The earliest versions of the TEI Guidelines were based on the Standard Generalized Markup Language (SGML). Since the publication of P4 in 2002, however, the guidelines have gradually been aligned with XML. Release P5 is entirely based on XML and the TEI Consortium notes that it is more dependent on emerging XML standards like schema languages and programming tools like XSLT and XQuery.[36]

The TEI Guidelines are not just concerned with metadata. As a 'markup language,' they are a means of defining XML tags that can both record metadata about a text - e.g., bibliographic description, provenance, or annotations - as well as enabling the encoding of the structural features of texts, including paragraphs, headings, titles, quotations, etc. Because the guidelines have been designed to work with a wide range of different text types, the full TEI tag set is quite large. For example, the TEI Consortium's website compares TEI P4's 500 (or so) tags with the 400 available in DocBook and the 90 in XHTML. However, it is noted that in practice, "most TEI users routinely use a much smaller subset of the full language."[37] For example, many adopters use a limited customisation of the TEI tagset known as *TEI Lite*, which the editors describe as being designed to meet "90% of the needs of 90% of the TEI user community" (Burnard & Sperberg-McQueen, 2006). *TEI Lite* emerged from the practical needs of the Oxford Text Archive and other electronic text centres, and earlier versions of the guidelines are available in number of different languages, including Chinese, French, Italian, and Spanish. The current version of *TEI Lite* is based on the TEI P5 release and is fully integrated with XML standards.

In addition, the TEI Consortium has worked with the Digital Library Federation to develop a sector-specific customisation of TEI called *TEI Tite*. These are intended to support libraries working with keyboarding vendors and are built upon guidelines produced by major US research libraries, chiefly the digital library production services at the universities of Michigan and Virginia, and the California Digital Library (Trolard, 2009). The main role of TEI Tite is to simplify the procurement of digitisation services and thus to reduce costs: "TEI developed TEI Tite to allow many of its smaller members and scholarly projects to procure digitisation services according to a standardized schema in a coordinated, discounted fashion."[38]

---

[35]TEI Consortium homepage; 2011: http://www.tei-c.org/index.xml Retrieved 13.03.2011
[36]TEI Consortium, *Introducing the Guidelines*; 2011: http://www.tei-c.org/Support/Learn/intro.xml Retrieved 13.03.2011
[37]*Ibid.*
[38]TEI Consortium, *TEI Tite digitization benefit: Request for proposals*, 2009: http://www.tei-c.org/Admin/RFP.xml Retrieved 13.03.2011

```
<?xml version="1.0" encoding="UTF-8" ?>
- <TEI xmlns="http://www.tei-c.org/ns/1.0">
  - <teiHeader>
    - <fileDesc>
      - <titleStmt>
          <title type="uniform">Gammer Gurton's needle</title>
          <title type="main">Gammer Gvrtons nedle [Electronic resource]</title>
        - <respStmt>
            <resp>possible author</resp>
            <name>Still, John, 1543?-1608</name>
          </respStmt>
        - <respStmt>
            <resp>possible author</resp>
            <name>Bridges, John, d. 1618</name>
          </respStmt>
        - <respStmt>
            <resp>possible author</resp>
            <name>Stevenson, William, d. 1575</name>
          </respStmt>
        - <respStmt>
            <resp>creation of machine-readable version</resp>
            <name>Lancashire, Ian</name>
          </respStmt>
        </titleStmt>
      - <extent>
          <seg type="designation">Text data</seg>
          <seg type="size">(1 file : ca. 103 kilobytes)</seg>
        </extent>
      - <publicationStmt>
        - <authority>
```

Example of part of TEI header

Example of part of TEI header. Source: Oxford Text Archive. Original text retrieved 19 January 2010 from: http://ota.ahds.ac.uk/headers/1788.xml

The TEI Header

A key part of the structure of any TEI document is a 'TEI Header' that will contain additional information (or metadata) about the encoded text. This would normally include descriptive metadata about the text or its original source - e.g., similar to the information included in a library catalogue record - but might also include some additional information on digitisation processes and encoding practice. An example of some of the types of information that can be contained within TEI header can be found in Figure 2. In practice, the descriptive information encoded within a TEI header will often overlap with metadata held in the other standards used by libraries, e.g. MARC (Machine-Readable Cataloging), MODS (Metadata Object Description Schema) or Dublin Core. While these may not always have a one-to-one correspondence, in many cases it should be possible to automatically generate some parts of TEI headers from these other records (and *vice versa*).

An integrated TEI-based resource: The Web interface of: Gwaith Dafydd ap Gwilym. Source: http://www.dafyddapgwilym.net/ Accessed 19.01.2010

An integrated TEI-based resource: The Web interface of: Gwaith Dafydd ap Gwilym. Source: http://www.dafyddapgwilym.net/ Accessed 19.01.2010

TEI in use

TEI was designed for the encoding of machine-readable texts to support research and education in the humanities, social sciences and linguistics. Since their first publication in 1994, the TEI Guidelines have been used by research and cultural-heritage organisations to create versions of texts that can be used for research and teaching. These include linguistic corpora as well as scholarly editions of historically significant texts (e.g. Burnard, O'Keefe & Unsworth, 2006). Prominent examples of literary and linguistic projects using TEI include the 100 million-word *British National Corpus* (e.g. Burnard, 2002; Kennedy, 2007) and scholarly editions of texts like the *Opere di Dante lemmatizzate* project[39] or the *Foxe's Book of Martyrs Variorum Edition Online*.[40] An example of the public interface of a online scholarly edition using TEI - in this case from the Welsh language *Gwaith Dafydd ap Gwilym* - can be found in Figure 3.[41] In this, the full-text of each poem can be viewed together with scholarly notes, information about manuscript sources (including transcriptions and images), and translations into modern language versions. Using TEI to generate these kinds of resources can be extremely resource-intensive.

TEI encoding levels in libraries

Specific guidelines for the use of TEI by libraries have been developed with the support of the Digital Library Federation (DLF), the TEI Consortium and other organisations. The current versions of these guidelines (Digital Library Federation, 2006; TEI Consortium, 2009) define five levels of encoding (Table 3), clarifying the distinct needs of different project-types. The levels defined range from fully automated projects that use TEI simply as a means of linking page images with raw-OCR output, to resource-intensive projects that implement a rich range of encoding from TEI P5 to create scholarly texts that are able to support detailed secondary analysis:

- Level 1: Fully Automated Conversion and Encoding

- Level 2: Minimal Encoding

- Level 3: Simple Analysis

- Level 4: Basic Content Analysis

- Level 5: Scholarly Encoding Projects

Moving up each level progressively enables more structural analysis of the text. Levels 1 and 2 are specifically reserved for those projects where the text being generated remains mainly subordinate to the page image and are probably the most suitable for the majority of mass-digitisation projects. Level 1 uses a very small number of tags; including page numbers for structure (and thus able to link with the page images), and with all of the OCR generated text wrapped within a single <ab> tag. Level 2 includes an additional means of identifying structural hierarchy in order to improve navigation through, for example, the generation of tables of contents, or similar. From Level 3 upwards, more attention is spent on the identification of logical structure, and such texts are intended to provide a foundation for upgrading texts to higher levels of encoding. The current version of the guidelines (v. 3) comments: "Level 3 generally requires some human editing, but the features to be encoded are determined by the logical structure and appearance of the text and not specialized content analysis" (TEI Consortium, 2009).

Using this scheme, the majority of mass digitisation projects using TEI would, therefore, aim to be encoded at levels 1 or 2.

Level 1	The text is generated through OCR, is subordinate to the page image, and is not intended to stand alone as an electronic text (without page images)

---

[39]University of Pisa, *Opere di Dante lemmatizzate;* University of Pisa; 2010: http://dante.di.unipi.it/ricerca/dante.html Retrieved 13.03.2011
[40]Foxe's Book of Martyrs Variorum Edition Online, v. 1.1; University of Sheffield; 2006: http://www.hrionline.ac.uk/johnfoxe/ Retrieved 13.03.2011
[41]Gwaith Dafydd ap Gwilym; University of Swansea; 2011:http://www.dafyddapgwilym.net/ Retrieved 13.03.2011

Level 2	The text is generated through OCR and is mainly subordinate to the page image, though navigational markers (textual divisions, headings) are captured
Level 3	The text is created by conversion from an electronic source such as HTML or word-processor documents or from a print source, either by way of OCR or keyboarding
Level 4	The text is generated either through corrected OCR or keyboarding and is able to stand alone without page images in order for them to be read by students, scholars, and general readers
Level 5	The text is generated either through corrected OCR or keyboarding and is able to stand alone without page images, as in Level 4. In addition, the tagging requires substantial human interventions by encoders with subject knowledge.

Encoding levels defined in: Best Practices for TEI in Libraries. Source: Retrieved 19 January 2010 from: http://wiki.tei-c.org/index.php/TEI_in_Libraries:_Guidelines_for_Best_Practices [http://wiki.tei-c.org/index.php/TEI_in_Libraries/_Guidelines_for_Best_Practices]

More information on the TEI

TEI Consortium website: Retrieved 13.03.2011 from: http://www.tei-c.org/Guidelines/access.xml

This is the most authoritative site for information on the TEI; provides access to the TEI Guidelines themselves, information on projects using the TEI, as well as a lot of supporting information (tutorials, membership details, etc.)

Best practices for TEI in Libraries: Retrieved 13.03.2011 from: http://wiki.tei-c.org/index.php/TEI_in_Libraries:_Guidelines_for_Best_Practices [http://wiki.tei-c.org/index.php/TEI_in_Libraries/_Guidelines_for_Best_Practices]

A document first produced in 1998 by a Digital Library Federation task force and updated several times since. This draft version of the guidelines is currently being updated by the TEI Special Interest Group on Libraries. It defines five levels of encoding to be used by digitisation projects, the first two of which are most suitable for mass-digitisation

Alternatives for packaging content and metadata

Standards like METS are primarily a means of packaging content and metadata into logical objects. While the packaging standard that has been most widely used in digitisation projects using OCR is a combination of METS with ALTO, a range of alternative content packaging frameworks exist. These include standards like the MPEG-21 Digital Item Declaration Language used in the aDORe repository (Bekaert, Hochstenbach & Van de Sompel, 2003) and more recently in European projects like DARE[42] and NEEO.[43] The following sections will provide a brief overview of two of recently developed content packaging standards: the EPUB format and the OAI-ORE specification.
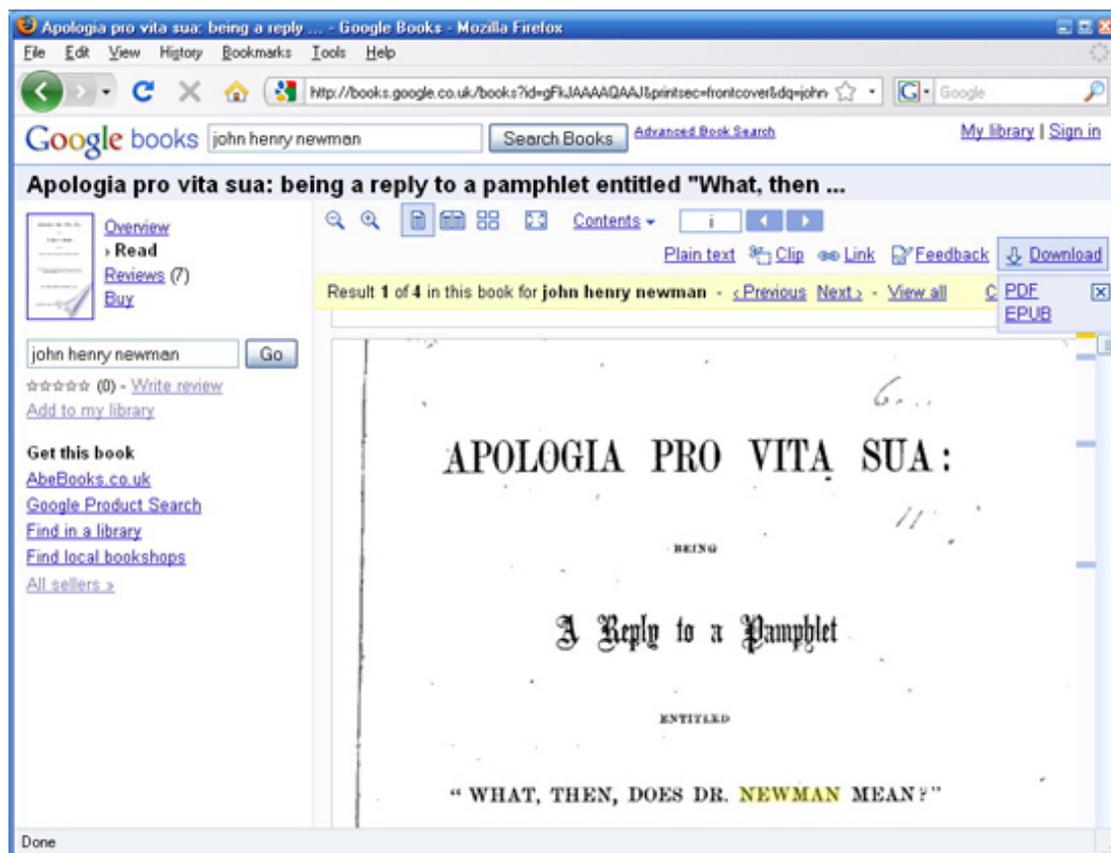
EPUB

EPUB is an open standard for the XML encoding of publications, produced by the International Digital Publishing Forum (formerly the Open eBook Forum). It superseded the earlier Open eBook Publication Structure (OEBPS) format. EPUB is one of a large number of XML-based formats primarily intended for use by eBook readers. The main advantage over primarily image-based delivery formats (like PDF)

---

[42]MPEG21 DIDL Application Profile for Institutional Repositories, v. 3.0; SURF Foundation; 2009: http://www.surffoundation.nl/wiki/display/standards/MPEG21+DIDL+Application+Profile+for+Institutional+Repositories [http://www.surffoundation.nl/wiki/display/standards/MPEG21%2BDIDL%2BApplication%2BProfile%2Bfor%2BInstitutional%2BRepositories] Retrieved 13.03.2011
[43]*NEEO Technical Guidelines*. NEEO Project Deliverable D5.3; 2008: http://www.neeoproject.eu/NEEO_TechGuide_0808.pdf Retrieved 13.03.2011

is that EPUB can adjust fonts for display on mobile devices, a concept known in the eBook reader world as "reflowable." Because the EPUB format is mainly based on text, it also facilitates annotation, font-size adjustments, and full-text searching. While a large number of eBook reader formats exists, EPUB is gradually beginning to look like emerging as a *de facto* industry standard, especially after Sony announced in August 2009 that it would convert its eBook store to the format.[44] In addition, the format has the potential to be used much more widely than for eBook readers. The development of browser plugins, like the EPUBReader for Firefox, will enable EPUB files to be opened and used within Web browsers.[45]

Google Books Web interface, showing options to download in both PDF and EPUB formats. Source: http://books.google.co.uk/

Google Books Web interface, showing options to download in both PDF and EPUB formats. Source: http://books.google.co.uk/

EPUB is composed of three open standards, the Open Publication Structure (OPS), the Open Packaging Format (OPF) and the Open Container Format (OCF). OPS and OPF define ways for representing all of the structure and markup relating to publications, while OCF provides the means (using ZIP files) of encapsulating this – together with alternative delivery formats, if required – into a single logical container that can be used for transmission, delivery and archival storage. OPF defines the mechanisms "by which the various components of an OPS publication are tied together and provides additional structure and semantics."[46] It does this through two XML files. The first defines the publications *metadata*, then provides a list of all files (the *manifest*), the linear reading order of content documents (the *spine*), and the

---

[44]*Sony converts eBook Store to EPUB format*; Sony Press Release; 2009: http://news.sel.sony.com/en/press_room/consumer/computer_peripheral/ e_book/release/41343.html Retrieved 21.03.2011
[45]EPUBReader; 2009: http://www.epubread.com/en/ Retrieved 21.03.2011
[46]*Open Packaging Format (OPF) 2.0 v1.0, Recommended Specification*; International Digital Publishing Forum; 2007: http:// www.openebook.org/2007/opf/OPF_2.0_final_spec.html Retrieved 13.03.2011

identification of particular structural elements (the *guide*). The second provides for navigation through a hierarchical table of contents.

Internet Archive Text Archive Web interface, showing options to download in full-text, PDF and DjVu; betas for EPUB, Daisy and Kindle formats. Source: http://www.archive.org/ Retrieved 19.01.2010

Internet Archive Text Archive Web interface, showing options to download in full-text, PDF and DjVu; betas for EPUB, Daisy and Kindle formats. Source: http://www.archive.org/ Retrieved 19.01.2010

Since August 2009, over a million out-of-copyright works digitised as part of the Google Books initiative have been made available in EPUB format. The EPUB files, like other versions of the text, are generated from the Google's OCR results. The broad process is described by Brandon Badger of Google.[47]

Google borrows the book from one of our library partners, much like you can from your local library. Before returning the book in undamaged form, we take photographs of the pages. Those images are then stitched together and processed in order to create a digital version of the classic book. This includes the difficult task of performing Optical Character Recognition on the page image in order to extract a text layer we can transform into HTML, or other text-based file formats like EPUB ...

The Internet Archive Text Archive also makes digitised texts available in EPUB and a number of other eBook reader formats (Figure 5). The EPUB format, therefore, provides a potential alternative delivery format for digitised texts. At the very least, digitisation projects creating METS or TEI-based files will probably need to explore whether they would need to create additional conversion routines to support the EPUB format.

---

[47]*"Download over a million public domain books from Google Books in the open EPUB format";* Badger, B.; 2009: http://booksearch.blogspot.com/2009/08/download-over-million-public-domain.html Retrieved 13.03.2011

Open Archives Initiative Object Reuse and Exchange (OAI-ORE)

The Open Archives Initiative (OAI) Object Reuse and Exchange framework is a recent attempt to provide a data exchange model for compound objects (aggregations) located on the Web. Like the OAI-PMH, it is an interoperability initiative of the Open Archives Initiative. The editors of the specifications describe a number of potential use cases. [48]

These standards provide the foundation for applications and services that can visualise, preserve, transfer, summarize, and improve access to the aggregations that people use in their daily Web interaction: including multiple page Web documents, multiple format documents in institutional repositories, scholarly data sets, and online photo and music collections.

OAI-ORE is primarily focused on fixing a particular problem with the Web architecture, i.e. that there is no standard mechanism that enables the identification or description of an aggregation of resources on the Web, posing significant challenges for machine-based applications, like search engines. In a white paper, Lagoze and Van de Sompel (2007) consider the case of a scanned book where all pages are made available on the Web and have been assigned HTTP URIs:

A crawler traversing the web may land on a resource corresponding to any page of the book, without regard to the actual order of the pages. Depending on the publishing approach taken by the information system, the crawler may obtain from this resource a representation that contains links to other scanned pages of the same book, or to the containing chapter or book. The representation may also contain links to related resources that are not part of the book, for example to resources that provide information about the author, the publisher, or to resources that are annotations, etc. Unfortunately these links are often un-typed, or if they do have type information, that type information cannot be automatically "understood" by the crawler since there is currently no standardized expression of link semantics on the web. Thus, a crawler or search engine cannot distinguish between the different links and notably which of the inter-linked resources correspond to the book and which do not.

In response to this problem, the OAI-ORE specifications provides a standardised way of describing the "constituents or boundary of an aggregation" on the Web (Lagoze and Van de Sompel, 2008). The specifications define a new class of (conceptual) Web resource known as *Aggregations*, which can be assigned URIs just like any other Web resource. Information about Aggregations is then made available through another resource known as a *Resource Map* (ReM). This also has a URI and provides information about the Aggregation in machine-readable form. McDonough (2009, p.2) notes the clear distinction made in the specifications between the abstract Aggregation and the Resource Map, "a concrete document that provides a serialized description of the aggregation." Resource Maps can be expressed in a variety of formats, including RDF/XML, RDFa and Atom XML.

OAI-ORE is not yet widely deployed in digitisation contexts. Its focus on the Web architecture means, however, that it may have an important future role where digitised content is made openly available on the Web. For example, researchers at the University of Illinois at Urbana-Champaign proposed experimenting with using OAI-ORE for the scholarly annotation of digitised content, including volumes digitised by the Open Content Alliance and other initiatives.[49] There has also been a certain amount of work working out how best to align OAI-ORE with METS (e.g., Habing & Cole, 2009; McDonough, 2009).

---

[48] *"Open Archives Initiative announces production release of Object Reuse and Exchange specifications"*; Open Archives Initiative; 2008: http://www.openarchives.org/ore/documents/ore-production-press-release.pdf
[49] *"Using OAI-ORE Resource Maps to support scholarly annotation of digitized books." Proposal from UIUC to the Andrew W. Mellon Foundation*; T. W. Cole, et al. 2008: http://oreo.grainger.uiuc.edu/docs/ColeOAI-OREProposalToMellonNoBudget-Jan08.pdf Retrieved 13.03.2011

Conclusions

This guide has attempted to provide a general overview of some of the areas where large-scale text digitisation projects need to consider metadata. It has attempted to define and outline the main categories of metadata and provide information on popular standards like METS and the TEI guidelines.

Things that may need to be considered in future guidance documents might include:

- **Automation -** linking metadata generation and capture with the industrial-scale processes and workflows typically used for large-scale digitisation

- **Quality control –** metadata quality is an extremely important issue. Geoffrey Nunberg (2009) has recently (and memorably) described Google Books as a "metadata train wreck", noting multiple errors in its descriptive metadata"[50] Google's Jon Orwant responded by noting that with over a trillion individual metadata fields, Google Book Search had *millions* of errors.[51] Digitisation at a large-scale amplifies any errors inherent in the source metadata, the imaging and OCR workflows, etc. The implication of this is that managing the quality of metadata will be an ongoing task for digitisation programmes.

- **Annotation layers –** Digitising texts is just the first step in a wider process that will involve the secondary reuse of content. This might include text enhancement, collaborative correction and the scholarly annotation or linking of text and metadata. The Australian Newspapers Digitisation Program has provided a prominent example of a project using the wider public to contribute corrections to OCR errors in a large database (Holley, 2009). The IMPACT project itself is exploring the use of linguistic and semantic tools for the enhancement of digitised content, including the use of historical dictionaries in OCR and the automatic identification of named entities.

- **Sharing metadata –** there will remain the need for digitisation programmes to share information with other services, e.g. registries of digitisation masters or something like the Book Rights Registry (BRR) established by the Google Book Search settlement. The main roles of these registries will be the management of an enormous amount of metadata. The Executive Director of BRR has said that its success will depend on the effective management of the large, complex and volatile metadata linked to Google's collection of seven to ten million digitised books.[52]

References

ANSI/NISO Z39.87-2006. Data dictionary -- Technical metadata for digital still images; National Information Standards Organization Web: http://www.niso.org/kst/reports/standards/ Retrieved 13.03.2011

Baca, M. (2008). "Introduction." In: *Introduction to metadata*, v. 3.0, ed. M. Baca. Los Angeles, CA: Getty Research Institute: http://www.getty.edu/research/conducting_research/standards/intrometadata/ Retrieved 13.03.2011

Beedham, H., Missen, J., Palmer, M., and Ruusalepp, R. (2005). *Assessment of UKDA and TNA compliance with OAIS and METS standards*. Colchester: UK Data Archive: http://www.data-archive.ac.uk/randd/ oaismets.asp Retrieved 13.03.2011

Bekaert, J., De Kooning, E., and Van de Walle, R. (2005). "Packaging models for the storage and distribution of complex digital objects in archival information systems: a review of MPEG-21 DID principles." *Multimedia Systems*, 10(4), 286-301.

---

[50] *"Google Books: a metadata train wreck"* University of Pennsylvania Language Log blog; Nunberg, G; 2009: http://languagelog.ldc.upenn.edu/ nll/?p=1701 Retrieved 13.03.2011

[51] *Ibid.*

[52] "What to expect from the Book Rights Registry." Book Business, 31 July 2009); Fletcher, H.: http://www.bookbusinessmag.com/article/what-expect-from-book-rights-registry-a-qanda-with-new-executive-director-michael-healy-410587_2.html Retrieved 13.03.2011

Burnard, L. (2000). "Text encoding for interchange: a new consortium." *Adriane*, 24: http://www.ariadne.ac.uk/issue24/tei/ Retrieved 13.03.2011

Burnard, L. (2002). "Where did we go wrong? A retrospective look at the British National Corpus." In: *Teaching and learning by doing corpus analysis*, ed. B. Kettemann & G. Marko (pp. 51-70). Amsterdam: Rodopi.

Burnard, L., Sperberg-McQueen, C. M. (2006). *TEI Lite: Encoding for interchange: An introduction to the TEI -- revised for TEI P5 release*: http://www.tei-c.org/release/doc/tei-p5-exemplars/html/teilite.doc.html Retrieved 13.03.2011

Clement, T., Steger, S., Unsworth, J., Uszkalo, K. (2008). *How not to read a million books*: http://www3.isrl.uiuc.edu/~unsworth/hownot2read.html [http://www3.isrl.uiuc.edu/%7Eunsworth/hownot2read.html] Retrieved 13.03.2011

Concordia, C. (2009). "Integration of heterogeneous metadata in Europeana." LIDA Workshop: Issues and Challenges with the Implementation of Metadata Schemes, Zadar, Croatia. http://dublincore.org/groups/tools/docs/LIDA09WorkshopC_1.pdf Retrieved 13.03.2011

Day, M. (2005). "Metadata." In: *DCC Digital Curation Manual*, ed. S. Ross & M. Day. Glasgow: Digital Curation Centre: http://www.dcc.ac.uk/resource/curation-manual/chapters/metadata/ Retrieved 13.03.2011

Fletcher, H. (2009). "What to expect from the Book Rights Registry: A Q&A with the new Executive Director, Michael Healey." *Book Business*, 31 July. http://www.bookbusinessmag.com/article/what-expect-from-book-rights-registry-a-qanda-with-new-executive-director-michael-healy-410587_2.html Retrieved 13.03.2011

Gartner, R. (2002). *Metadata Encoding and Transmission Standard (METS)*. JISC Techwatch Report TSW 02-05. http://www.jisc.ac.uk/whatwedo/services/techwatch/reports/horizonscanning/hs0205.aspx Retrieved 13.03.2011

Gartner, R. (2003). *MODS: Metadata Object Description Schema*. JISC Techwatch Report TSW 03-06: http://www.jisc.ac.uk/whatwedo/services/techwatch/reports/horizonscanning/hs0306.aspx Retrieved 13.03.2011

Gilliland, A. (2008). "Setting the scene." In: *Introduction to metadata*, v. 3.0, ed. M. Baca. Los Angeles, CA: Getty Research Institute. http://www.getty.edu/research/conducting_research/standards/intrometadata/ Retrieved 13.03.2011

Habing, T., and Cole, T. (2009). *Candidate approaches for describing ORE Aggregations in METS: preliminary discussion draft*. University of Illinois at Urbana-Champaign. http://ratri.grainger.uiuc.edu/oremets/ Retrieved 13.03.2011

Hakala, J. (2001). Using National Bibliography Numbers as Uniform Resource Names. RFC 3188. Internet Engineering Task Force: http://tools.ietf.org/html/rfc3188 Retrieved 13.03.2011

Hilse, H.-W, and Kothe, J. (2006). *Implementing persistent identifiers: Overview of concepts, guidelines and recommendations*. London: Consortium of European Research Libraries; Amsterdam: European Commission on Preservation and Access. http://www.knaw.nl/ecpa/publ/pdf/2732.pdf Retrieved 13.03.2011

Holley, R. (2009). "How good can it get? Analysing and improving OCR accuracy in large scale historic newspaper digitisation programs. *D-Lib Magazine*, 15(3/4), March/April. http://www.dlib.org/dlib/march09/holley/03holley.html Retrieved 13.03.2011

Hurley, B. J., Price-Wilkin, J., Proffitt, M., and Besser, H. (1999). *The Making of America II Testbed Project: a digital library service model*. Washington, D.C.: Council on Library and Information Resources. http://www.clir.org/pubs/abstract/pub87abst.html Retrieved 13.03.2011

ISO 15836:2009. Information and documentation – The Dublin Core metadata element set. Geneva: International Organization for Standardization.

Kennedy, G. (2007). "An under-exploited resource: using the BNC for exploring the nature of language learning." In: *Corpus Linguistics and the Web*, ed. M. Hundt, N. Nesselhauf, & C. Biewer (pp. 151-165). Amsterdam: Rodopi.

Klijn, E. (2008). "The current state-of-art in newspaper digitization." *D-Lib Magazine*, 14(1-2), January/February: http://www.dlib.org/dlib/january08/klijn/01klijn.html Retrieved 13.03.2011

Kunze, J. (2003). "Towards electronic persistence using ARK identifiers." http://www.cdlib.org/inside/diglib/ark/arkcdl.pdf Retrieved 13.03.2011

Lagoze, C., and Van de Sompel, H. (2007). *Compound information objects: the OAI-ORE perspective*. Open Archives Initiative White Paper. http://www.openarchives.org/ore/documents/CompoundObjects-200705.html Retrieved 13.03.2011

Lagoze, C., and Van de Sompel, H. (2008). *ORE User Guide – Primer*. Open Archives Initiative. http://www.openarchives.org/ore/1.0/primer Retrieved 13.03.2011

Lee, B. (2009*). Use of METS and ALTO in the Australian Newspapers Digitisation Program (ANDP) at the National Library of Australia (NLA)*, v. 2.0 (August). http://www.nla.gov.au/ndp/project_details/documents/ANDP_Use_of_METSv2.pdf Retrieved 13.03.2011

Littman, J. (2006). "A technical approach and distributed model for the validation of digital objects." D-Lib Magazine, 12(5), May. http://www.dlib.org/dlib/may06/littman/05littman.html Retrieved 13.03.2011

McDonough, J. P. (2009). "Aligning METS with the OAI-ORE Data Model," in Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries, Austin, TX, June 15-19, 2009. New York: Association for Computing Machinery. http://hdl.handle.net/2142/10744 Retrieved 13.03.2011

Nunberg, G. (2009). Google's Book Search: a disaster for scholars. *Chronicle of Higher Education*, August 31: http://chronicle.com/article/Googles-Book-Search-A/48245/ Retrieved 13.03.2011

Paradigm Project (2007). *Paradigm Workbook on Digital Private Papers: Persistent identifiers*. http://www.paradigm.ac.uk/workbook/metadata/pids.html Retrieved 13.03.2011

TEI Consortium. (2002). *TEI P4: Guidelines for electronic text encoding and interchange*, ed. C. M. Sperberg-McQueen and L. Burnard. http://www.tei-c.org/release/doc/tei-p4-doc/html/ Retrieved 13.03.2011

Best practice guides

General introductions to metadata:

Baca, M. (ed.) (2008). *Introduction to metadata*, v. 3.0. Los Angeles, CA: Getty Research Institute. http://www.getty.edu/research/conducting_research/standards/intrometadata/ Retrieved 13.03.2011

Caplan, P. (2003). *Metadata fundamentals for all librarians*. Chicago, IL: ALA Editions.

Greenberg, J. (2005). "Understanding metadata and metadata schemes." *Cataloging & Classification Quarterly* 40(3/4), 17-36: http://www.ils.unc.edu/mrc/pdf/greenberg05understanding.pdf Retrieved 13.03.2011

National Information Standards Organization. (2004). *Understanding metadata*: http://www.niso.org/publications/press/UnderstandingMetadata.pdf Retrieved 13.03.2011

JISC Digital Media (2009) Metadata. http://www.jiscdigitalmedia.ac.uk/crossmedia/advice/metadata-standards-and-interoperability/ Retrieved 13.03.2011

Digitisation:

Federal Agencies Digitization Guidelines Initiative: http://www.digitizationguidelines.gov/ Retrieved 13.03.2011

Oxford Digital Library, *Metadata in the Oxford Digital Library*. http://www.odl.ox.ac.uk/metadata.htm Retrieved 13.03.2011

Minerva, *Digitisation Guidelines: a selected list*: http://www.minervaeurope.org/guidelines.htm Retrieved 13.03.2011

Puglia, S., Reed, J., and Rhodes, E. (2004). *Technical guidelines for digitizing archival materials for electronic access: Creation of production master files - raster images*. College Park, MD: National Archives and Records Administration. http://www.archives.gov/preservation/technical/guidelines.html Retrieved 13.03.2011

National Library of Australia, Digitisation guidelines. http://www.nla.gov.au/digital/standards.html Retrieved 13.03.2011

National Library of New Zealand, Digitisation guidelines. http://www.natlib.govt.nz/catalogues/library-documents/digitisation-guidelines/ Retrieved 13.03.2011

Yale University Library Digital Production & Integration Program, *Best Practices*: http://www.library.yale.edu/dpip/bestpractices/ Retrieved 13.03.2011

Identifiers:

Hilse, H.-W, and Kothe, J. (2006). *Implementing persistent identifiers: Overview of concepts, guidelines and recommendations*. London: Consortium of European Research Libraries; Amsterdam: European Commission on Preservation and Access: http://www.knaw.nl/ecpa/publ/pdf/2732.pdf Retrieved 13.03.2011

Paradigm Project (2007). *Paradigm Workbook on Digital Private Papers: Persistent identifiers*. http://www.paradigm.ac.uk/workbook/metadata/pids.html Retrieved 13.03.2011