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 **ix**
Preface **xi**
Acknowledgements **xiii**

PART I METADATA CONCEPTS **1**

1 Introduction **3**
Overview 3
Why metadata? 3
Fundamental principles of metadata 4
Purposes of metadata 11
Why is metadata important? 17
Organisation of the book 17

2 Defining, describing and expressing metadata **19**
Overview 19
Defining metadata 19
XML schemas 24
Databases of metadata 26
Examples of metadata in use 27
Conclusion 33

3 Data modelling **35**
Overview 35
Metadata models 35
Unified Modelling Language (UML) 36
Resource Description Framework (RDF) 36
Dublin Core 39
The Library Reference Model (LRM) and the development of RDA 40
ABC ontology and the semantic web 42
Indecs – Modelling book trade data 44

OAIS – Online exchange of data 46
Conclusion 48

4 Metadata standards 49
Overview 49
The nature of metadata standards 49
About standards 51
Dublin Core – a general-purpose standard 51
Metadata standards in library and information work 54
Social media 62
Non-textual materials 64
Complex objects 70
Conclusion 74

PART II PURPOSES OF METADATA 75

5 Resource identification and description (Purpose 1) 77
Overview 77
How do you identify a resource? 77
Identifiers 78
RFIDs and identification 85
Describing resources 86
Descriptive metadata 88
Conclusion 93

6 Retrieving information (Purpose 2) 95
Overview 95
The role of metadata in information retrieval 95
Information Theory 97
Types of information retrieval 98
Evaluating retrieval performance 102
Retrieval on the internet 104
Subject indexing and retrieval 106
Metadata and computational models of retrieval 107
Conclusion 111

7 Managing information resources (Purpose 3) 113
Overview 113
Information lifecycles 113
Create or ingest 117
Preserve and store 118
Distribute and use 122
Review and dispose 123
Transform 124
Conclusion 124

8 Managing intellectual property rights (Purpose 4) 127
Overview 127
Rights management 127

Provenance 134
Conclusion 137

9 Supporting e-commerce and e-government (Purpose 5) 139
Overview 139
Electronic transactions 139
E-commerce 140
Online behavioural advertising 141
Indecs and ONIX 143
Publishing and the book trade 144
E-government 148
Conclusion 149

10 Information governance (Purpose 6) 151
Overview 151
Governance and risk 151
Information governance 153
Compliance (freedom of information and data protection) 154
E-discovery (legal admissibility) 156
Information risk, information security and disaster recovery 156
Sectoral compliance 158
Conclusion 159

PART III MANAGING METADATA 161

11 Managing metadata 163
Overview 163
Metadata is an information resource 163
Workflow and metadata lifecycle 164
Project approach 165
Application profiles 170
Interoperability of metadata 171
Quality considerations 179
Metadata security 181
Conclusion 182

12 Taxonomies and encoding schemes 185
Overview 185
Role of taxonomies in metadata 185
Encoding and maintenance of controlled vocabularies 186
Thesauri and taxonomies 188
Content rules – authority files 191
Ontologies 194
Social tagging and folksonomies 199
Conclusion 201

13 Very large data collections 203
Overview 203
The move towards big data 203

What is big data? 205
The role of linked data in open data repositories 206
Data in an organisational context 209
Social media, web transactions and online behavioural advertising 211
Research data collections 212
Conclusion 219

14 Politics and ethics of metadata 221
Overview 221
Ethics 221
Power 226
Money 229
Re-examining the purposes of metadata 230
Managing metadata itself 236
Conclusion 237

References 239
Index 257

List of figures and tables

Figures

1.1	Metadata from the Library of Congress home page	12
2.1	Example of marked-up text	20
2.2	Rendered text	21
2.3	Word document metadata	28
2.4	Westminster Libraries – catalogue search	30
2.5	Westminster Libraries catalogue record	30
2.6	WorldCat search	31
2.7	WorldCat detailed record	32
2.8	OpenDOAR search of repositories	32
2.9	Detailed OpenDOAR record	33
3.1	An RDF triple	37
3.2	More complex RDF triple	37
3.3	A triple expressed as linked data	38
3.4	DCMI resource model	39
3.5	Relationships between Work, Expression, Manifestation and Item	41
3.6	LRM agent relationships	42
3.7	Publication details using the ABC Ontology	44
3.8	Indecs model	45
3.9	OAIS simple model	46
3.10	OAIS Information Package	46
3.11	Relationship between Information Packages in OAIS	47
4.1	BIBFRAME 2.0 model	57
4.2	Overlap between image metadata formats	66
4.3	IIIF object	67
4.4	Relationships between IIIF objects	67
4.5	Metadata into an institutional repository	72
4.6	How OAI-PMH works	72
5.1	Example of relationship between ISTC and ISBN	85
5.2	Structure of an Archival Resource Key	85

6.1 Resolution power of keywords 96
6.2 Boolean operators 100
6.3 British Library search interface 108
6.4 Metadata fields in iStockphoto 111
7.1 DCC simplified information lifecycle 116
7.2 Generic model of information lifecycle 116
7.3 PREMIS data model 121
7.4 Loan record from Westminster Public Libraries 123
8.1 ODRL Foundation Model 131
8.2 Legal view of entities in ONIX 132
8.3 Creative Commons Licence 133
8.4 PROV metadata model for provenance 135
9.1 Cookie activity during a browsing session 142
9.2 ONIX e-commerce transactions 146
11.1 Stages in the lifecycle of a metadata project 166
11.2 Singapore Framework 170
11.3 Possible crosswalks between four schemas 177
11.4 Possible crosswalks between ten schemas 177
11.5 Data Catalog Vocabulary Data Model 178
11.6 A-Core Model 180
12.1 Extract from an authority file from the Library of Congress 192
12.2 Conceptual model for authority data 192
12.3 Use of terms from a thesaurus 193
12.4 Google Knowledge Graph results 197
12.5 Structured data in Google about the British Museum 198
13.1 Screenshot of search results from the European Data Portal 208
13.2 Agents involved in delivering online ads to users 212
13.3 A 'pyramid' of requirements for reusable data 214
13.4 Silo-based searching 218
13.5 Federated search service 218
13.6 Index-based discovery system 219

Tables

1.1 Day's model of metadata purposes 13
1.2 Different types of metadata and their functions 14
4.1 KBART fields 60
4.2 IIIF resource structure 68
11.1 Dublin Core to MODS Crosswalk 176
13.1 Comparison of metadata fields required for data sets in Project Open Data 209
13.2 Core metadata elements to be provided by content providers 213
14.1 Metadata standards development 231

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.

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)