1

Definitions and scope

This chapter defines the book's scope and aims, and introduces key concepts that will be discussed at much greater length in the following chapters. The book's title and subtitle are explained first, and then how information resource description fits into the modern information environment. This is followed by an outline of the field of information organization and an overview of the book's structure.

Information resources

This book is about the description of information resources. Just about all types of resource can be described, and are. However, this book focuses on *information* resources. We are not so much interested in how vacuum cleaners, for instance, are described by sales people; we are more interested in, for example, how books (including e-books) are described by librarians. On the other hand, we *are* interested in how vacuum cleaners are described by museum curators: as museum objects, vacuum cleaners can provide us with information about, for instance, their mechanical development. Thus, just about everything *can* be an information resource, depending on the context, and so we need to avoid defining information resources too narrowly. We are interested here in the description of all resources that, in the context of their description, are primarily intended to inform.

The word 'primarily' in that last sentence is worth including. A very sophisticated vacuum cleaner could, say, inform its user when the dust bag is full, but, as a cleaning tool, the vacuum cleaner's primary function is to clean, not to inform. As a museum exhibit, however, its primary function would indeed be to inform. Conversely, many resources found in museums, libraries, archives and other 'information agencies' (they are sometimes also

referred to as 'memory institutions') may do more than just inform, even in those contexts (they may also entertain, for instance), but a primary function is nevertheless to provide the patron with information.

Ultimately, resources are described as information resources according to the view of the describer, even if they have not been created as an information resource and are not generally used as one. However, those resources that are created primarily to inform and that are mostly used for the information they contain are those most likely to be described as information resources, and so will be given greater coverage. Thus, readers may be relieved to learn, we will talk more about resources such as books than about resources such as vacuum cleaners.

Resources have been created for the purpose of providing information for a very long time. They are essentially communication devices, designed to disseminate messages. Not all communication devices are information resources, however. Many devices, such as the human vocal chords, produce 'live' messages, for a particular moment in time. We are more interested in those devices that contain messages, re-transmittable across time. As such, these resources carry *recorded* information. The information may come in a variety of forms, but whatever its form, it is, at least potentially, re-accessible and re-usable. In today's world of digital information, this functionality may sometimes be taken for granted, but it has underpinned the development of all human civilizations. Many of the inventions that have advanced the recording of information – writing, printing, photography, computers and so on – have had profound effects on human history. Indeed, they have, in a very literal way, made history.

Sometimes recorded information is referred to as recorded *knowledge*; similarly, information resources such as books could be considered and are sometimes called 'knowledge resources'. Indeed, many books in the field of information organization (also referred to as knowledge organization) use the terms 'information' and 'knowledge' interchangeably. Knowledge managers sometimes refer to the pyramid set out in Figure 1.1, with *wisdom* at the top and *data* at the bottom, but for our purposes it might be more useful to view this as a continuum based on the amount of processing carried out in the generation of a message's content. Data may represent very little processing, in which, for example, observations are directly recorded from a science experiment; information may involve an analysis of data; knowledge may represent an integration of different pieces of information and a conclusion; and wisdom may represent a reflection on this conclusion, in the

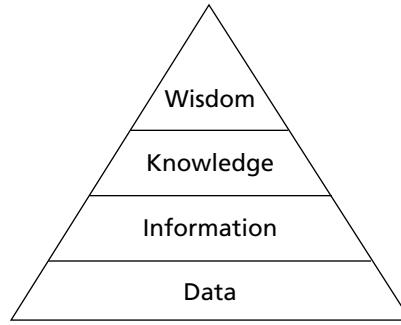

Figure 1.1 *The 'knowledge pyramid'*

light of other conclusions and experience. As 're-accessible message resources' is a bit of a mouthful, this book will use the term 'information resource' as the generic term to cover all those resources that contain, or represent, data, information, knowledge and/or wisdom. For one thing, 'information' conveys a communicative aspect (to inform) that is not present in the other three terms.

Resource description

We now turn to the other half of the book's title. As we have noted, all kinds of resources can be described, for all kinds of different purposes. Information resources are described, ultimately, for the purpose of facilitating the use of the information they contain. The reasons why describers want this information to be used vary. Authors may want to become famous, publishers may want sales, librarians may want to improve their patrons' knowledge. The various reasons need to be borne in mind, as they may well have a bearing on the nature of the description.

The questions of how and what description might facilitate the use of information resources are addressed throughout the book, but we will make a start here. First, we need to recognize that, as a communication process, resource description involves both describers and recipients. Describers need not be directly associated with the resource, though sometimes they are (as in the case of authors). Describers need not even be human: computers are able to automatically generate descriptions for all kinds of digital information resource. While different describers may have different roles to play, as well as different motives, they almost always intend their descriptions to have recipients. These

recipients may be a particular group of people or the population at large or (at least in the first instance) computers. They may even be the describers themselves, as in the case of individuals compiling personal reading lists. The intended audience of a description, as well as the describer, is likely to influence its nature, or at least should do, if the description is to be effective.

Just as our focus is on recorded information, so too is it on recorded description (of that information). The medium in which the description is recorded is also significant, of course. Descriptions recorded on sheets of paper or on index cards are retrieved and used in different ways than are those recorded in online databases. Indeed, the nature of the description may well vary according to the kind of system, not just the medium, in which it is stored. For example, a description is more likely to include certain data elements if the system indexes those elements.

Essentially, descriptions consist of information about different aspects of the thing they describe. We have just referred to these as *data elements*. In the case of information resources, elements may pertain to the information itself (the *content*) or to the *carrier* of this information (all recorded information is dependent on some sort of carrier, such as a book, a roll of film or a computer file) or to both content and carrier. Examples of content elements include subject and language; examples of carrier elements include size and physical location (if there is one).

Information resource descriptions come in all shapes and forms. They may be long or short, containing many elements or very few. Recorded descriptions tend to be textual, or at least verbal, but do not have to be. This book, however, concentrates on those descriptions or elements of description that are primarily used to (effectively) access information resources, as explained in the next section.

Metadata

As a product, information resource description is quite often referred to as *metadata*, which literally means data about data (more definitions are provided by Liu, 2007, 5; see also Greenberg, 2005). We have already given ourselves licence to use the term 'information' for data, and we will likewise use the term 'metadata' as another term for 'information resource description' (and treat it as a singular noun). In this vein, metadata covers data elements that pertain to the carriers of information, as well as those that pertain to the information (content) itself.

We should note, however, that the term 'metadata' is often associated specifically with *digital* information resources and is commonly defined as 'structured' data (about data), which in this context means data that can be processed by computer. In most cases, the metadata covered in this book will operate in the electronic environment and thereby conform, at some level, to a certain structure. However, even the non-electronic metadata dealt with here is likely also to be structured, so that it can be used in some form of non-computerized information retrieval system. For instance, card indexes are structured, with headings at the top and various other data elements set out underneath.

It should also be noted that 'metadata' is sometimes associated with other kinds of resource, apart from information resources, such as other products and services transacted through e-commerce. These do not directly concern us, although they may well influence the ways in which information resources are described, particularly in cross-domain applications.

Various categories of metadata, specifically for information resources, have been identified by different writers. Often metadata is characterized by its function. For example, Haynes' 'five-point model of metadata' (2004, 17) is based on five purposes: resource description; information retrieval; resource management; ownership and authenticity; and interoperability. While all metadata might ultimately be intended to facilitate the use of resources, there is a range of ways in which this occurs. For instance, resources may need to be managed so that they remain usable, hence Haynes' category of 'resource management' metadata. This kind of metadata might be needed to inform, for example, the preservation of a resource. Another term for this category is *administrative* metadata (though sometimes preservation metadata is distinguished from other forms of administrative metadata). Resources, particularly digital resources, may also need to be assembled and rendered for (human) use. The metadata used to facilitate this is often called *structural* metadata. Last, but by no means least, information resources also need to be accessed, i.e. retrieved. In Haynes' model, 'retrieval performance' is supported by 'resource description' and 'information retrieval' metadata. It is the metadata that supports the provision of access to, and retrieval of, information resources that is the focus of this book.

Metadata that facilitates access to information resources is sometimes referred to as 'discovery metadata', although access is a broader concept than discovery. Another term used is 'descriptive metadata', even though all

metadata is ultimately descriptive (Caplan, 2003). For our purposes, the generic term will suffice: 'metadata' will be used to mean this specific kind of metadata, unless otherwise indicated. 'Information resource description' will likewise be used in this narrower sense, as it has been in the book's title.

Metadata can support effective access to information resources in several ways. As well as indicating to the prospective user how to obtain them, metadata can help the user decide whether they should be obtained. Moreover, metadata is often used to advise the user of their existence in the first place. It can also be used to provide an overview of a collection of resources by grouping like resources together (otherwise known as *collocation*), allowing the user to 'navigate' it. Similarly, metadata can be used to navigate a single resource, or to facilitate access to particular components of a resource. All of these functions will be investigated.

The metadata discussed in this book is mostly intended for use in *information retrieval systems* of various kinds, designed to facilitate access to collections of information resources (or resource components). As we have noted, these systems constrain the nature of the metadata, including its structure. We shall pay particular attention to those systems developed by information agencies such as libraries and archives, which have been a leading force in the field for a very long time. We shall deal with *personal* information retrieval systems only to the extent that their nature coincides with those intended for wider use.

Finally, it might also be pointed out that 'metadata' is by no means the only other term that can be used for information resource description. Terms such as 'cataloguing', 'bibliographic data', 'indexing', 'archival description' and 'museum documentation' are commonly employed in particular contexts, and will be in this book too.

Elements, values, format and transmission

An information resource can be looked at, and described, in all sorts of ways. There are, in fact, an infinite number of possible *metadata elements*. It may, of course, be more useful to record some attributes of a resource than others. The weight of a book, for instance, might not be particularly helpful for the purposes of resource discovery (people generally don't search for books of a particular weight, unless they are after a door stop); its title, on the other hand, may well be (people often search for books by title). Chapter 2 introduces some of the more commonly recorded elements.

Producing effective metadata, however, is not just about choosing the right elements. It is also about using appropriate *values* to record these elements. Most commonly, these values comprise words, such as the words of a title. Sometimes they comprise numbers (e.g. an International Standard Book Number), or other kinds of representation. Just as there can be any number of possible metadata elements, there can also be many options when it comes to recording their values. For example, the name of an author might be recorded as 'Joe Bloggs', or 'Joseph Bloggs', or 'J. H. Bloggs' etc., while the subject of a resource could be, say, 'Animals' or 'Fauna'. Some values are likely to be more effective than others, and the values used often have a large impact on the quality of metadata, discussed at length in Chapter 5. They need to be accurate, of course (unless one is looking to mislead), but they also need to possess certain other qualities, such as intelligibility.

Further, these values need to be recorded in an appropriate *format.* Importantly, the format needs to be compatible with the information retrieval system for which the metadata is intended. The metadata may also need to be input into the system in a particular way, i.e. using a particular *transmission protocol.*

These different aspects of metadata creation can be illustrated by the analogy of bottles being labelled, filled, stored and delivered (Elings, 2007). The elements of description provide a *structure*: they are the bottles. A catalogue record, comprising various *fields* (one for the title, another for the name of the author, another for the name of the publisher etc.), is an example of such a structure. These bottles are then filled with particular values, i.e. *content.* Note that not all bottles need be filled for a particular description: some fields may not be applicable for particular resources (a resource might not necessarily have an International Standard Book Number, for instance).

Once the bottles have been filled, they need to be stored in boxes. Similarly, the metadata needs to be *encoded* in a particular format. This can be for the benefit of people and/or machines, so that they can read and process the metadata. An example of an encoding scheme is HTML, a mark-up language that can be processed by web browsers. Finally, the boxes may be delivered to a particular place (although they could, alternatively, be taken out of storage and consumed directly). Packages of metadata, typically in the form of computer files, are often fed into other systems through certain protocols that they have been programmed to act upon.

All four aspects of metadata – elements, values, format and transmission – are examined in this book, in some cases simultaneously, as they are not

always so readily distinguishable. What is important to note from the outset, however, is that different elements, values, formats and protocols are effective in different information contexts, depending on the characteristics of the users, the technology, the information resources and other environmental factors.

Managing metadata

The subtitle of this book indicates that it covers not only the process of describing information resources, i.e. creating metadata, but also that of managing the metadata once created. There are various ways in which metadata is managed, an activity typically carried out by information professionals.

First, metadata may be obtained from an external source, rather than being created in-house. As we have already noted, not all metadata is equal, and the quality of metadata from different sources, or even the same source, may need to be evaluated.

Second, metadata may need to be fed into an information retrieval system, which may entail converting it into a compatible format and applying certain protocols.

Third, metadata may need to be improved by editing or adding to it. As a process, this is similar to metadata creation, but is even more likely to be done with reference to particular users, systems and costs.

Fourth, metadata needs to be presented to users as effectively as possible. Modern computing enables systems to provide a wide range of user interfaces; different interfaces may suit different users and search contexts.

Fifth, metadata needs to be preserved, if it is to be reused. Rapid technological change has created greater, rather than less, risk to the integrity of data. All too quickly, data can be 'lost', due to the obsolescence of both software and hardware. In today's digital environment, systems and metadata cannot afford to be left behind.

Sixth, the significant costs involved in metadata creation and management can often be reduced through exchange mechanisms, which means giving as well as receiving. When providing metadata to other agencies, it may be necessary to comply with external policies and standards: the efficient sharing of metadata usually requires a degree of standardization, and sometimes data conversion, so that the metadata can be used effectively in other systems. Standardization is a key concern of many metadata managers.

All these areas of metadata management require a detailed understanding of the functions of metadata, in relation to particular user contexts, just as the creation of metadata does. Creating and managing metadata can sometimes be a rather technical business, but decisions and policies should ultimately be made with reference to what works for the end-user.

The contemporary information environment

We have outlined the scope of this book by discussing its title and subtitle. It is important to put information resource description, as a process and product, into context. Metadata is not created and managed in isolation. We have already alluded to some of this context – describers, users, systems and so forth. Clearly, with so many variables affecting information resource description, its relationship with the wider information environment is complex and ever changing. This does not mean, however, that metadata is arbitrary; on the contrary, it is frequently conventionalized and standardized. For example, the nature and format of the metadata presented in and on information resources has evolved into conventions such as title pages, film credits and record labels.

Metadata works only if people understand and use it. Thus, it could be argued that all effective metadata is based on certain social conventions, such as language. When conventions are deemed important enough, they sometimes become *standards*, which are essentially formal agreements to adhere to certain practices. The convention of the title page, for instance, may become a standard if a particular community considers it to be an essential component of a text. Thus, universities require students to include title pages (set out in a particular way) in the theses and dissertations they submit. Over the past century or so, library cataloguing conventions have become increasingly standardized, to the extent that most cataloguers in the English-speaking world now apply the same set of content rules and use the same record structure and encoding system.

Standardization may encourage particular practices, but it does not prevent practices from changing. Just as laws change according to evolving social norms, so metadata standards change according to the evolving information environment. New types of information resource require new conventions and standards. For example, websites do not have title pages, but instead have home pages. Metadata standards for existing resources may also change as new technologies become available. For instance, in the first

half of the 20th century, library catalogue records were produced on standard 3 by 5 inch cards. By the end of the century, this standard had become all but obsolete – instead, the vast majority of records were being created according to standards for electronic records. Metadata creation and management, even when conventionalized and standardized, is never insulated from changes in information technology and information behaviour; rather, it is always an integral part of an information *ecology*, in which information agents constantly interact with their information environment.

The contemporary information environment encompasses an extraordinary range of technologies. Older technologies are still widely used: many people still read printed books, newspapers and magazines; historians still visit physical archives; people still hang paintings on their walls. However, the digital technologies developed over the past few decades have transformed the information environment beyond all recognition, both qualitatively and quantitatively. Just about any form of information can now be recorded digitally, and not just by a privileged few but by the empowered many. Personal computers, let alone corporate servers, can now store huge amounts of data, while anyone with an internet connection can disseminate information to millions of people across the world, often in a matter of seconds. Mobile technologies are further increasing the pervasiveness and utility of the online information environment.

Although there are still large parts of the world's population with little or no access to digital technology, it is becoming ever more affordable. In the developed world, creating and using digital information has become an everyday activity for a majority of people. As a result, the amount of recorded information has increased at a mind-boggling rate, with very significant implications for the task of information retrieval.

This information revolution has given rise to information resources with very different characteristics from those based on older technologies. According to Liu (2004), digital resources exhibit greater information density; less longevity; less uniqueness; greater duplicability; greater mobility and fluidity; greater connectivity; and greater integration. Online information resources tend to be far less permanent than their analogue counterparts. Their longevity may depend on just one computer server: if it is disconnected, or a file on it is deleted, a resource may be lost to the world for ever. Digital resources can also be readily manipulated, so that new versions can be made with ease; they can likewise be duplicated at the touch of a button. Information resources published on the web are connected to each other in a

very powerful way, through the hyperlink. This enhances both physical and intellectual access to documents, and to components of documents.

These characteristics raise various issues when it comes to the description of digital information resources. Should each version of a website be described? Should the various components of the website be described as resources in their own right? Indeed, who should do the description? Furthermore, which, and how many, of the millions of information resources now available online should be described? Different approaches to information resource description are necessary in the new information world in which we find ourselves.

Digital technologies have already had a profound effect on metadata creation and management. This book examines the new approaches to information resource description brought about by these technologies, but also recognizes the value and continuing validity, in certain contexts, of traditional practices. The digital information revolution does not, of course, constitute a complete break with the past, and it is likely that older technologies and practices will continue to influence the way we describe information resources for a long time to come. Even in the web environment, we still talk of 'pages', for instance. However, it needs to be acknowledged that digital information is not only here to stay, but will become even more pervasive in the future. The book considers the description of all information resources, both physical and online, from this perspective.

Information organization

The practice of information resource description is central to the field commonly known as *information organization* or *knowledge organization.* As such, it is part of the broader field of information management. The term 'organization' is used here to refer to various ways in which information resources are organized so as to improve access to them, and not just to their *physical arrangement*. Clearly, how objects are arranged in physical space has a bearing on the extent to which they may be accessed effectively. The same can be said for how links to digital resources are arranged on a web page. However, when there are a lot of resources and/or a lot of space (either physical or virtual), other organizing techniques may need to be utilized. This applies to resources of all kinds, not only information resources.

In a kitchen, for example, all the spice jars might be placed together on a particular shelf. This arrangement is based on a categorization (spices)

intended to narrow down the search for a particular spice. Instead of having to remember where each foodstuff is kept in a random arrangement, the cook just has to remember where the spices are kept.

However, the cook still has to identify which of the spice jars has the particular spice they need. The cook could inspect the contents of the various jars, but it might be easier if the jars were labelled. Again, the aim is to facilitate access to the desired ingredient. *Labelling* is also a common way of organizing information resources. The spines of books, for example, are typically labelled with titles, author names and, in a library, call numbers. Likewise, links to web resources are usually labelled with words that indicate their content. All such labels are metadata, since they describe information resources.

Appropriate labelling, in addition to convenient arrangements, is likely to make for a more organized kitchen, so that the cook can find the things they want more easily. However, this may still not be enough. Instead of a kitchen, suppose the cook is looking for a particular spice in a large supermarket. Perhaps it is shelved in a general 'spices' section, or amongst the 'Asian foods', or in some other section, or perhaps the supermarket does not stock the particular spice. All the jars in the supermarket may be sensibly arranged and clearly labelled, but the cook may not be familiar with the supermarket and may end up spending a lot of time walking up and down the aisles in search of the elusive spice. It might be better to consult a database of all the stock available in the store.

Most databases comprise one or more *indexes.* In the supermarket's database, the spice might be indexed (if it is stocked) under its generic name and brand name. Typically, the user merely types in a name and the computer does the rest, searching the index and displaying the details for the item, including its location in the store. Essentially, an index consists of *representations* of the actual resources, arranged in a way that makes them easier to search through than the resources themselves. In the case of indexes to information resources, these representations may be metadata, such as titles and author names, or they may be derived from the information content itself, such as the words found in a text. They may be viewed as 'conceptual' labels, and the index as an arrangement of labels, combining the two basic organizing techniques.

If indexing can be useful in the supermarket context, it is often vital for the effective use of libraries, archives and other collections of information resources. Most of the tools developed to provide access to these collections,

such as library catalogues, archival finding aids, museum registers and search engines, are essentially indexes. Many are also based on metadata. The content of individual information resources, such as books, serials and maps, is also often accessed via indexes.

Whether indexed or not, metadata is very much at the heart of information organization practice. However, we should recognize too the importance of *content-based* indexes in modern information retrieval. Without doubt, these indexes, epitomized by search engines such as Google, cover a far greater number of information resources than do the traditional tools based on metadata, and in many situations are very effective. Although this book does not cover content-based information retrieval per se (it is generally regarded as a separate field), its continuing advance is reflected upon in later discussion concerning the value and future of metadata-based information retrieval.

The field of information organization deals with information packaged up as resources, which are also commonly referred to as *documents*, even when they contain non-textual information. Information organization is thus concerned with *document retrieval.* Most information retrieval is essentially document retrieval of one kind or another, although it also covers efforts to provide information that answers particular questions much more directly than would a whole document, or even a section of a document. These efforts are on-going and represent an interesting area of research, but are beyond the scope of this book.

There are a couple more terms that need to be explained in relation to the field of information organization. *Bibliographic control* and *bibliographic organization* generally refer to the practice of information organization in libraries. Although the term 'bibliographic' literally pertains to books, librarians tend to use it in a more general sense, for all information resources provided by libraries. A primary mode of bibliographic control is library cataloguing. However, although some books on information organization concentrate on this activity, it should already be apparent that it is by no means the only mode of information resource description, and this book aims to cover the field much more broadly. In recent times, many writers have distinguished the traditional bibliographic approach embodied in library cataloguing from newer forms of information resource description represented by the term 'metadata'. There are certainly differences, but, as Howarth (2005, 37) argues, there are also sufficient similarities 'to warrant a confluence in terminology and definition'.

Overview of this book

The next chapter discusses the nature of metadata in terms of its likely elements. The discussion is based on a consideration of why metadata is needed and used. The third chapter then looks at an important part of the context surrounding its use, namely, the various systems and tools dedicated to information retrieval. Chapter 4 considers another critical contextual aspect, the creators of the metadata. We then turn our attention back to the users: how does information resource description best serve their needs? Chapter 5, therefore, looks at metadata quality. Chapter 6 returns the discussion to the systems utilizing metadata and examines how they go about sharing metadata. Standardization is a key quality of effective metadata, as well as of system interoperability, and metadata standards are surveyed at length in Chapters 7 and 8. First, standards that focus on elements, format and transmission are examined, and then the vocabularies used for values. The final chapter considers the future of information resource description through a discussion of the prospects for the different approaches to information retrieval. A list of selected resources for further reading is provided at the end of the book, followed by a list of the metadata standards covered.

References

Caplan, P. (2003) *Metadata Fundamentals for all Librarians*, American Library Association.

Elings, M. W. (2007) Metadata for All: descriptive standards and metadata sharing across cultural heritage communities, *VRA Bulletin*, **34** (1), 7–14, www.vraweb.org/seiweb/readings-prep/MetadataforAll_Elings-Waibel.pdf.

Greenberg, J. (2005) Understanding Metadata and Metadata Schemes, *Cataloging & Classification Quarterly*, **40** (3/4), 17–36.

Haynes, D. (2004) *Metadata for Information Management and Retrieval*, Facet Publishing.

Howarth, L. (2005) Metadata and Bibliographic Control, *Cataloging & Classification Quarterly*, **40** (3), 37–56.

Liu, J. (2007) *Metadata and its Applications in the Digital Library: approaches and practices*, Libraries Unlimited.

Liu, Z. (2004) The Evolution of Documents and its Impacts, *Journal of Documentation*, **60** (3), 279–88.