Practical Ontologies for Information Professionals

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

Practical Ontologies for Information Professionals

David Stuart

facet publishing

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

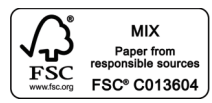
Facet Publishing is wholly owned by CILIP: the Chartered Institute
of Library and Information Professionals.

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library.

ISBN 978-1-78330-062-4 (paperback)
ISBN 978-1-78330-104-1 (hardback)
ISBN 978-1-78330-152-2 (e-book)

First published 2016

Text printed on FSC accredited material.



Typeset from author's files in 10/13 pt Minion Pro and Myriad Pro by
Facet Publishing Production.
Printed and made in Great Britain by CPI Group (UK) Ltd, Croydon, CR0 4YY.

Contents

List of figures and tables

Figures

Tables

CHAPTER 1

What is an ontology?

Introduction

Today more data and information are being produced and shared than ever before; data is streaming forth from new online social behaviours as well as high-specification digital tools and instruments. If we are to extract the maximum value from this data then we need to make use of the most appropriate tools and technologies. Ontologies, formal representations of knowledge with rich semantic relationships, are one such tool, and the focus of this book.

This chapter provides an introduction to ontologies, and considers their increasing importance to information professionals. Following a brief overview of the growing information overload and data deluge, the chapter considers the various definitions that have been applied to the term 'ontology' and how ontologies differ from associated and overlapping information concepts such as controlled vocabularies, taxonomies, metadata and knowledge bases. Finally, the chapter considers the potential of ontologies for information retrieval and discovering 'undiscovered public knowledge', and the role of the librarian in the development, maintenance and curation of ontologies.

The data deluge and information overload

It is important to start with an understanding of the changing information landscape, reminding ourselves of why we need new tools and technologies, and why it is no longer acceptable to continue with the way things have always been done. We are awash with a wide variety of information and data, but due to the tools that we are currently using the value of much of the data is going to waste. As John Naisbitt (1984, 17) put it, 'We are drowning in information, but starved for knowledge'.

Information is coming from a wide variety of sources. There has been an explosion in the publishing and sharing of text across the whole of the communication spectrum, from the informal to the formal. Traditional formal publications, such as books and journals, have been joined by e-books and e-journals, with new publishing models based on combinations of self-publishing and open access: the number of self-published

titles published in the USA rose from 85,468 titles in 2008 to 458,564 titles in 2013 (Bowker, 2014); whilst Chen (2014) estimated that the proportion of articles published in the previous year available as open access had either passed or was very close to 50%.

In the middle of the formal–informal spectrum of publishing is the grey literature: white papers, reports, technical papers and other, more informal, publications. Whereas once this grey literature could be costly to create and had limited circulation, desktop publishing software and electronic publishing on the web have put it within reach of a wide range of individuals and organizations. But the growth in these numbers has been dwarfed by the growth of social media and other informal publishing, where the associated numbers are often in the hundreds of millions if not billions: there are 1.49 billion active Facebook users each month (Facebook, 2015); and over 500 million updates are sent on Twitter on a typical day (Twitter Engineering Blog, 2013). No one can hope to read anything but the smallest fraction of this information, even within the smallest of fields. There is a need for new tools to help with information retrieval, increasing precision without excessively impacting recall.

The narrative text has also been joined by increasing quantities of other text, such as computer code and data sets, as well as rich media (i.e., images and video). Although the lack of data sharing within the academic community has been labelled as the 'dirty little secret' of open science data promotion (Borgman, 2012, 1059), the potential of open data and open code to transform the rate of scientific progress (Hey, Tansley and Tolle, 2009) and to encourage more open and accountable governments and encourage citizens' participation (Raman, 2012) has led to numerous open programs and policies. Governments have signed up to open data charters promising data to be open by default (Cabinet Office, 2013) and funding agencies and journals are increasingly stipulating the need for open data and open code (e.g., *Nature*, 2014). It is not enough, however, that data and code are open; they need to be findable and reusable by those who want to make use of them too.

Whilst the growth of open data may have been slower than some would like, growth in the number of images and videos shared has exploded: since its launch in 2010, over 30 billion images have been shared on Instagram (Instagram, 2015); in May 2014 Snapchat reported 700 million photos sent per day (Techcrunch, 2014); and YouTube counts billions of views every day as people watch hundreds of millions of hours of video (YouTube, 2015). This media is also increasingly of higher quality, part of the trend towards increasingly high specification digital tools and instruments. By 2007 83% of mobile phone cameras had digital cameras, and over the years the specification of these cameras has increased dramatically. By 2012 there were mobile phones with 41 megapixel cameras available, many times more powerful than the first camera phones with 0.1–1 megapixels. The rise of increasingly high specification mobile phone cameras reflects an increase in digital data collection at increasingly high-level

specifications across a wide range of disciplines and professions. Data per 360 degree scan in computed tomography has gone from 57.6 kB in 1972 to 0.1–1GB by 2010 (Kalender, 2011), whilst the rise in quality and fall in price has increased the number of scans made and the areas outside medicine where computed tomography may be used (e.g., archaeology and paleontology). When the first human genome was declared complete in 2003 it had been a mammoth project taking over ten years and costing US$3billion; now we have entered the US$1000 genome era, where the cost of sequencing the human genome has fallen to a price where it may play a role in predictive and personalized medicine (Hayden, 2014). Projects such as the 100,000 Genome Project are now sequencing thousands of genomes to identify genetic causes for a wide range of human diseases (www.genomicsengland.co.uk/the-100000-genomes-project).The content in any single human genome, however, is dwarfed by the amount of data produced by big science projects such as the Large Hadron Collider, where 19 gigabytes of data were created in the first minute and thirteen petabytes (10^{15} bytes) in the first year (Brumfiel, 2011). With so much data available, and in increasingly large chunks, it becomes increasingly important that we are accessing and downloading only the most relevant data for analysis.

As well as the data people are making a conscious decision to share, there are also the vast digital trails we all increasingly leave as an increasing proportion of our lives are lived online, and processes are digitized. Mobile phones can not only capture pictures, but have built in GPS and accelerometers to track location and movement. Phone (or VOIP) calls can now simply be captured in their entirety, to index or playback in full at a later date if necessary. With the internet as the first port of call for our information needs we are leaving trails of information about the searches we are carrying out, the pages we are visiting and the links we are following. This information is not only restricted to the log files of a single site, but may be aggregated by advertising companies and content providers across multiple sites, enabling the building of increasingly complex profiles on individuals for the tailoring of increasingly personalized advertising and services.

As data storage and processing prices have fallen it is no longer necessary to be selective in what we capture: increasingly we capture everything and then search the captured information for what we need later. A process that is epitomized by note-taking software designed for capturing 'everything' and ideas such as life streaming. Wearable technology, such as Google Glass, streamlines the process, as it is no longer necessary to even go to the trouble of taking a smartphone from a pocket.

Data inevitably produces more data. The data that is captured is often indexed, analysed, or combined to spawn more data. A file may be indexed, the contents analysed according to different criteria (e.g., searching for patterns or antecedents), and be accompanied by an ever growing quantity of descriptive, access, and preservation

metadata. As new questions are asked, and new methods of data analysis developed, the same data set can continue to produce ever increasing quantities of data. We have entered the era of Big Data. There are vast amounts of structured and unstructured data available, and there are new challenges to ensure that we make use of this data.

Neither the exponential growth of science nor the problems of information overload are particularly new problems. The growth and communication of science began to be explored scientifically in the 1950s and 60s, and its exponential growth was one of the subjects of Derek J. de Solla Price's (1963) seminal *Little Science, Big Science*. The history of scientific publishing can be seen as one of trying to help researchers overcome the problem of information overload, first with publication of specialist journals, then with specialist abstract and indexing services. However, the web has provided a step-change in the publishing of information. When Ziman (1969) wrote of the problem of having to wade through 'tomes of irresponsible nonsense' without peer review, he would have had no idea how large these tomes of irresponsible nonsense would become.

The web requires new tools and methods to help users engage with the information that is available, and its brief history has already been one of rapid innovation: from directories to search engines, from information searching to information discovery. We no longer expect always to have to search for the information that we require, but are instead alerted to information we may require, either through the filter of social network sites or algorithmic suggestions (e.g., Google Scholar).

Those who successfully find ways of managing the information overload, and of making use of the increasing quantities of data available, will have the competitive advantage. Whether that is the company gathering competitive intelligence on its rivals, the researcher looking for new ways to encode and analyse data, or the international non-governmental organization looking for efficiencies in sharing information.

Ontologies are one way of helping to tame some of the problems identified above, providing a structure for this information in such a manner that it can be read automatically and unambiguously, and shared more widely.

Defining terms

Whenever writing on a specialist subject it is generally advisable to start by defining your terms, as all too often we follow the example of Humpty Dumpty when he says in Lewis Carroll's *Through the Looking Glass*: 'When I use a word, it means just what I choose it to mean – neither more nor less'. Even within the smallest of fields the same term may have multiple meanings, some of which may be conflicting, a feature that is true for both 'ontology' and concepts such as data, information and knowledge, which the ontology is trying to encode.

Defining data, information, knowledge and wisdom

Most topics in information science can't be discussed for long without running into the terms data, information, or knowledge. Unfortunately the terms are notoriously hard to define, and attempts at capturing knowledge within the library and information science community (e.g., through knowledge management) have sometimes been controversial for seemingly being little more than rebranding exercises.

Data, information, knowledge and wisdom are often conceptualized as a four-step pyramid, from data at the bottom, through information and knowledge, to wisdom at the top. This model was popularized by Ackoff (1989), but analysis of how the terms are used (Rowley, 2007; Zins, 2007) finds them to be the subject of wide-ranging and often overlapping definitions. Rather than thinking of them as distinct terms, it is more useful to think of them as overlapping areas on a continuum from highly structured and codified information at one end (data) to highly personal tacit understanding at the other (wisdom).

Data is the 'building blocks' of information and knowledge (Kitchin, 2014), although much of the information and knowledge that we have can seem quite detached from the underlying data. Whereas the route from data to knowledge may seem quite direct in the hard sciences, within the arts and the humanities the relationships between abstract ideas and concepts that form information and knowledge are less readily structured. Ontologies emerged as a way of capturing knowledge, and codifying it in a highly structured manner as data, and this may be applied to knowledge in any discipline.

> . . . knowledge is inherently complex and the task of capturing it is correspondingly complex. Thus, we cannot afford to waste whatever knowledge we do succeed in acquiring.
>
> Neches et al., 1991, 54

Knowledge organization systems and ontologies

Ontologies are one of a number of different knowledge organization systems that have been developed within the information profession to improve information discovery. These knowledge organization systems are also variously known as 'taxonomies' or 'controlled vocabularies', depending on the sector within which they are used. Whereas cultural heritage institutions err more towards 'controlled vocabularies', the commercial sector tends to use the term 'taxonomies'.

Harpring (2013, 13) defines a controlled vocabulary as: 'an organized arrangement of words and phrases used to index content and/or to retrieve content through browsing or searching', very similar to Hedden's broad definition of a taxonomy in her introduction to *The Accidental Taxonomist*:

> . . . any knowledge organization system (controlled vocabulary, synonym ring, thesaurus, hierarchical term tree, or ontology) used to support information/content findability, discovery, and access.
>
> Hedden, 2010, xxii

There is also a more narrow use of the term taxonomy, in the sense it refers to a hierarchical set of terms (Hedden, 2010; Harpring, 2013), such as the Linnaean taxonomy of biological classification, most people's first introduction to the term. Within this work the term controlled vocabulary is preferred rather than taxonomy, partly due to the potential for confusion caused by the dual meaning, but also due to the author's own background within library and information science.

Controlled vocabularies have both advantages and disadvantages. Advantages of a controlled vocabulary include improved recall and greater precision through reducing polysemy (van Hooland and Verborgh, 2014). Recall, the proportion of relevant documents that are retrieved out of all the relevant documents in a collection, is increased by the reduction of the number of terms associated with a particular concept. For example, the Dublin Core Metadata Initiative Type Vocabulary is a controlled vocabulary of 12 terms: collection, dataset, event, image (still image and moving image), interactive resource, physical object, service, software, sound, and text. Without a controlled vocabulary, a wide range of resources that adhere to each of these types could have been referred to differently. The 'text' resource type includes letters, books, theses, reports, newspapers, and poems, as well as a host of other texts primarily designed for reading. To ensure the recall of all the associated text resources would require entering all the possible terms.

Polysemy refers to multiple meanings for the same term. A controlled vocabulary enables distinctions to be made between the different terms. For example, 'Apple' may refer to the fruit, the technology company, a computer created by the technology company, or the record label founded by the Beatles. Within the Library of Congress Subject Headings the fruit has the term 'Apples' and the computer is 'Apple computer', whilst in the Library of Congress Name Authority File the technology company is 'Apple Computer, Inc.' and the record label is 'Apple Records'.

There are also a number of disadvantages to controlled vocabularies: the cost, the complexity, the slow evolution, and their subjectivity (van Hooland and Verborgh, 2014). Controlled vocabularies are not only expensive to create in the first place, but also to maintain as new names and terminology enter a field.

In some situations the slow speed of change may be simply due to limitations in resources; in other situations there may be conflict between the terminology of conservative and progressive perspectives. For example, a comparison of the style guides of left- and right-wing newspapers can be particularly enlightening regarding

their associated politics. Controlled vocabularies are inevitably subjective, and reflect the world view of the creators at a particular time, and different people in more enlightened times inevitably baulk at previous decisions, especially when there are prohibitively large legacy costs to rectifying previous decisions. For example, the Dewey Decimal Classification system is infamous for class 200 – religion, where seven out of the ten divisions relate to the Bible or Christianity:

- 200 Religion
- 210 Philosophy & theory of religion
- 220 The Bible
- 230 Christianity
- 240 Christian practice & observance
- 250 Christian pastoral practice & religious orders
- 260 Christian organization, social work, & worship
- 270 History of Christianity
- 280 Christian denominations
- 290 Other religions.

Although there have been attempts to extend many of the other religions in DDC in recent years, particularly Islam (Idrees, 2012), the Dewey legacy nonetheless supports the perception of it being Christian-centric.

Some of the most widely used forms of controlled vocabularies within the information profession are subject headings, authority files and thesauri. It is worth considering each of these types of controlled vocabulary, and their limited nature, for comparison with the more expressive nature of ontologies:

Subject headings are a controlled set of terms designed to describe the subject or topic of a resource, whether it is book, article, or data set. Popular examples include the Library of Congress Subject Headings (http://id.loc.gov/authorities/subjects.html) and the Medical Subject Headings (MeSH) (www.nlm.nih.gov/mesh/meshhome.html). Subject heading lists ensure that the same term is used to describe a work, rather than multiple similar terms.

Authority files are sets of preferred headings. As well as preferred subject headings, there may be preferred organization names, person names, and place names. History is replete with people, places, and organizations that have different names at different times, and successful information retrieval requires the consistent use of terms and relationships between the alternatives: those looking for information on Mark Twain may also want to retrieve information on Samuel Clemens, whilst those researching Constantinople may also wish to retrieve information on Istanbul. Well known examples include the authority files of the major national libraries (e.g., Library of Congress, British

Library and Bibliothèque Nationale de France). VIAF (Virtual International Authority File) (http://viaf.org) is a project from several national libraries designed to link together the separate authority files of the libraries into one virtual authority file.

A **thesaurus**, like a taxonomy (in the narrower sense of the term), provides hierarchical relationships between concepts (i.e., broader and narrower terms), as well as equivalence and associative relationships. A typical entry in a thesaurus might include all three types of relationship, as in the example below for information science:

```
Information Science
Broader terms:    Sciences
Narrower terms:   Computer Science
                  Library Science
Use instead of:   Informatics
                  Information Industry
Related terms:    Information Processing
                  Information Skills
                  Knowledge Management
                  Knowledge Representation
                  Library Education
```

The above example is based on 'Information Science' in the ERIC (Education Resources Information Center) thesaurus (http://eric.ed.gov). The relationships within a thesaurus enable a reader to traverse from one concept to another more easily, helping to find related content. Other well known examples of thesauri include the Getty Thesaurus of Geographic Names (www.getty.edu/research/tools/vocabularies/tgn), the Art & Architecture Thesaurus (www.getty.edu/research/tools/vocabularies/aat), and the Thesaurus for Graphic Materials (www.loc.gov/pictures/collection/tgm) from the Library of Congress.

Today controlled vocabularies should also be compared with tagging, which came to prominence with the rise of social media and social networking sites. The vast size and diversity of the web, and its users, drove the need for an approach to classification that was equally global and diverse in outlook, and could be applied by members of the public as well as information professionals. Tagging, the application of uncontrolled terms to online resources, has been incorporated into a large number of services with varying degrees of success. Whilst many of the sites for bookmarking web resources (e.g., del.icio.us) have fallen out of favour, it nonetheless continues to have an important role within sites that are focused around user-generated content: for example, the tagging of images in Flickr and Instagram, and the use of hashtags in Twitter (so called because of the '#' used to denote the tag). In comparison to a

controlled vocabulary, tagging is likely to have reduced recall and lack precision, but where the scale of the web is concerned there may be few alternative options.

An ontology is like a thesaurus, in that there are multiple types of relationship between terms, but it can be non-hierarchical, with a far richer set of relationships, and typically holds a far greater variety of information. The richness of the relationships and information means that it is not only suitable for indexing resources, but may be a knowledge base for knowledge discovery in its own right.

Defining an ontology

Ontologies first emerged in the Artificial Intelligence (AI) community, borrowing the term 'ontology' from philosophy, where ontology is concerned with the study of being or existence. The term was adopted by the AI community in the 1980s for computational models that can enable automated reasoning (Gruber, 2009), having recognized that 'capturing knowledge is the key to building large and powerful AI systems' (Neches et al., 1991, 37).

Today the most widely used definition of ontology is Gruber's (1993, 199) definition: 'an explicit specification of a conceptualization'. This has been criticized for its broadness, incorporating both simple glossaries and 'logical theories couched in predicate calculus' (Gruber, 2009, 1964), and also for its focus on subjective concepts rather than entities as they exist in reality (Smith, 2004). Nevertheless, an ontology might be considered a near-synonym with knowledge organization system or taxonomy (in the broad sense).This continuum from informal vocabularies to formal ontologies has been reiterated by the World Wide Web Consortium (W3C) in their introduction to ontologies: 'There is no clear division between what is referred to as "vocabularies" and "ontologies"' (W3C, 2013). The broadness of the definition is an important part of the inclusiveness of ontologies for information professionals. It is not just a subject for the AI community, but rather all those involved in the codifying of knowledge, including librarians, archivists, museum workers and domain experts. Nonetheless, a more specific definition is useful for distinguishing between those ontologies that are the primary focus of this book and other examples of controlled vocabularies.

Within most definitions of ontologies the distinctive feature of ontologies is the richness of the relationships between terms. For Hedden (2010, 12), an ontology 'can be considered a type of taxonomy with even more complex relationships between terms than in a thesaurus . . . it aims to describe a domain of knowledge, a subject area, by both its terms . . . and their relationships'. Within an ontology a person does not have to just be related to an event: they may be present at an event, organize an event, take part in an event, be an authority on an event, or possibly instigate an event.

An example of the richness of the information associated with a particular entity in an ontology is provided below with an author record:

Ranganathan, S.R. (Shiyali Ramamrita), 1892-1972

event:	1892
	1972
family name:	Ranganathan
given name:	S.R.
has created:	Colon classification / S.R. Ranganathan
	The five laws of library science / S.R. Ranganathan
name:	S.R. Ranganathan
type:	Agent
	Person
has contributed to:	An essay in personal bibliography / A.K. Das Gupta
same as:	49268668

The above record is based on the British National Bibliography record for S.R. Ranganathan. It expresses two types of relationship between the author and his associated works: has created, and has contributed to. With the exception of the name, family name, and given name values, each of the properties on this record links to another record for the particular instance, for example, *The five laws of library science*:

The five laws of library science / S.R. Ranganathan

bnb:	GB6417211
description:	2nd ed originally published (B58-927) Madras Library Association; Blunt 1958.
edition statement:	2nd ed. reprinted (with minor amendments)
type:	BibliographicResource
creator:	Ranganathan, S.R. (Shiyali Ramamrita), 1892-1972
is part of:	Ranganathan series in library science; no 12
language:	eng
publication event:	Asia Publishing House, 1964
same as:	GB6417211
subject:	020

Again, many of the properties have their own associated records, creating a huge graph

of related resources, joining previously disparate authority lists and classification systems. Figure 1.1 shows the graph produced by just the author and instance records mentioned above.

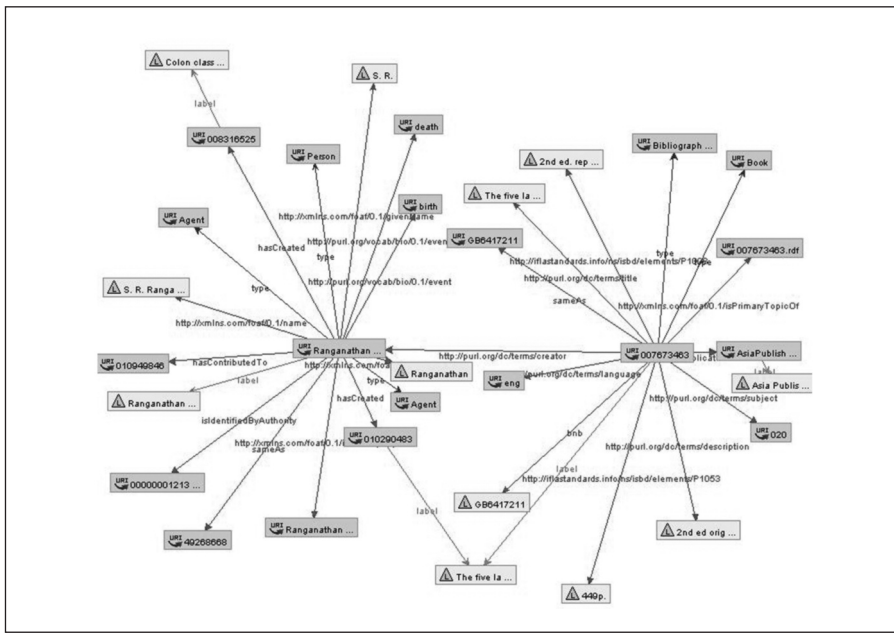


Figure 1.1 *Section of the British National Bibliography graph visualized using RDF Gravity*

Explicit specifications of conceptualizations are important if computers are to successfully communicate with one another without ambiguity, and there is less ambiguity and more scope for drawing inferences if the explicit specifications build upon one another in a more formal manner. 'Formal' rather than 'explicit' is used in a number of definitions of ontologies: 'An ontology is a formal specification of a shared conceptualization' (Borst, 1997,11); 'Ontologies are formalized vocabularies of terms, often covering a specific domain and shared by a community of users. They specify the definitions of terms by describing their relationships with other terms in the ontology' (W3C, 2012). Others, however, have preferred to combine the two terms: 'An ontology is a formal and explicit specification of a shared conceptualization' (Jakus et al., 2013, 29). Whilst a formal ontology would seem to necessitate an ontology being explicit, an explicit ontology does not necessarily need to be particularly formal. The use of relationships in defining terms is a particularly important part of the semantic web due to its distributed nature, with organizations likely to be adhering to different vocabularies.

As well as the richness of the relationships and their explicitness, there is another distinctive feature of ontologies that is widely acknowledged: that they should be a representation of the structure of knowledge, not just a set of indexing terms. Willer and Dunshire (2013, 112) define an ontology as 'a formal representation of the structure of knowledge and information, and Allemang and Hendler (2011, 1) point out that semantic models are sometimes called ontologies.

Although Harpring (2013) acknowledges certain similarities between thesauri and taxonomies and ontologies, she considers them to have fundamentally different goals:

> …ontologies use strict semantic relationships among terms and attributes with the goal of knowledge representation in machine-readable form, whereas thesauri provide tools for cataloguing and retrieval.
>
> Harpring, 2013, 26

The goals of knowledge representation and information retrieval do not have to be mutually exclusive, however, and the same ontology may be used for both. In fact the richness on the relationships may allow for far richer querying and information retrieval.

Within this book a fairly broad definition of ontology, albeit not quite as broad as that of Gruber (1993), is taken:

> An ontology is a formal representation of knowledge with rich semantic relationships between terms.

Such ontologies may be more or less formal, depending on the extent to which they define terms with relation to one another and incorporate axioms, and no distinction is made as to whether an ontology is designed either for information retrieval or as a knowledge base. Such a simple definition, however, glosses over the parts that comprise an ontology.

The parts of an ontology

The definition of an ontology provided above is designed to be inclusive, although it is sometimes necessary to distinguish between different ontologies that fall within this definition. As with Willer and Dunshire's (2013) definition, it is sometimes used to distinguish the structure of the ontology from the instances. For example, a book ontology might not be expected to include any information about particular books, but rather provide the necessary structure for describing books and the relationships between them and associated types of objects. In other situations an ontology might

refer to both the structure and the instances, in much the same way as a thesaurus of place names includes the names of places, not just the possible relationships between them.

Whether an ontology developer is interested primarily in classes or instances may be expected to differ considerably depending on the discipline. For example, Arp, Smith and Spear, who are primarily interested in the representation of scientific research, believe an ontology is 'concerned with representing universals' (2015, 17). However within the arts and humanities it may be the particular facts that are important rather than the general theories, and the general theories do not necessarily have widespread agreement.

The W3C Library Linked Data Incubator Group (2011) makes a distinction between metadata element sets and value vocabularies within data sets, with the metadata element set providing the structure for holding the information (e.g., Dublin Core element set) and the value vocabularies providing the values for these elements (e.g., an authority list of author names or place names). This book also distinguishes between the structure and the values of ontologies, although it uses slightly different terminology:

- ontology element set
- ontology instances.

The ontology element set and ontology instances combine to form an ontology data set or knowledge base.

The term *metadata* is one that is already overburdened within the information profession, and may cause confusion when distinguishing between more traditional approaches to cataloguing and the rich semantic nature of ontologies. Metadata is also strongly associated with a particular type of record within the information profession (e.g., a bibliographic record describing a book), and it is important that ontologies are more inclusive than this.

'Instances' is a more inclusive term than 'value vocabulary', which seems primarily appropriate for existing controlled vocabularies, whereas an instance may be used to refer to any concept or thing within an ontology. A concept is generally an abstract idea that is then given a label, some of which are more concrete than others (e.g., 'Paris' may be considered a more concrete concept than 'Love'), but which are nonetheless abstract. Concepts form the basis of most traditional knowledge organization systems, but ontologies can also deal with more concrete things. As well as the abstract idea of Paris, the one that each of us holds in our minds, with associations of romantic getaways, literary salons or fashion shows, there is the actual physical city with specific boundaries, activities and population at any particular moment. Concepts and things

are often blurred within ontologies, but there is nonetheless a wide range of information associated with any particular concept or thing that is not part of many controlled vocabularies. Following Hedden's (2010, 69) use of 'term record', *instance record* is used to describe all the pieces of information associated with a particular concept or thing, or *resource record* within the context of the semantic web.

For ease of reading, once an ontology element set or an ontology data set has been introduced as such in this book, the subsequent text may simply refer to it as an 'ontology' or either an 'element set' or a 'data set'. Ontologies differ greatly, but all represent a formal representation of knowledge with rich semantic relationships between terms.

Types of ontology

Just as we reach a point where the reader is likely to believe they have an understanding of what an ontology consists of, it is necessary to introduce a range of additional terminology that has been adopted to describe types of ontologies. Here we briefly describe four of them: lightweight ontologies; upper ontologies; application profiles; and ontology languages.

Usability is an important consideration when it comes to the creation of ontologies, but as Murdock, Buckner and Allen (2012) ask: '. . . usability by whom or by what?' Some have argued that ontologies are 'unsuited to the rough-and-tumble of real-world applications once they get beyond a certain level of complexity' (Brewster and O'Hara, 2007, 565). Lightweight ontologies are ontologies that are designed for ease of use, processable by machines but also accessible to humans, focusing on core classes (i.e., types of entities) and properties rather than constraints and axioms (Rocha da Silva et al., 2014). These may be particularly important in the humanities, where concepts are far less concrete or widely agreed upon. It is lightweight ontologies that have the widest use, especially on the semantic web, and are the type of many of the ontologies within this book.

An upper ontology (also known as a foundation ontology) is a general all-inclusive ontology that can theoretically connect all others. Such an ontology can aid ontology interoperability and alignment, and provide a starting point for developing more specific domain ontologies (Opalički and Lovrenčić, 2012). Examples of upper ontologies include Suggested Upper Merged Ontology (SUMO) (www.adampease.org/OP), OpenCyc (www.cyc.com/platform/opencyc) and the Basic Formal Ontology (http://ifomis.uni-saarland.de/bfo). Whether a single, universal ontology is feasible or desirable for representing the myriad of views and perspectives from different domains is open to debate, and is often ignored in the linked data approach to a semantic web. In this work the focus is less on upper ontologies, and more on what may be referred

to as middle-level ontologies, those that are not designed to be universal but are nonetheless designed to accommodate data from a large number of domains. These include Europeana Data Model and CIDOC-CRM, both of which are returned to in Chapter 3, along with one upper ontology, the Basic Formal Ontology.

Application profiles have been defined as: '. . . schemas which consist of data elements drawn from one or more namespaces, combined together by implementors, and optimized for a particular local application' (Heery and Patel, 2000). They reflect the practical application of ontologies to meet real-world needs that may differ considerably from strict standards described in the original documentation. Increasingly, however, attempts have been made to accommodate the differences in the requirements of the standard makers and the implementers. Dublin Core Terms were developed with application profiles and the semantic web in mind (Baker, 2012), whilst Resource Description and Access (RDA) has both constrained and unconstrained properties, with the unconstrained properties being independent of the overarching Functional Requirements for Bibliographic Records (FRBR) model and having no explicit range or domain. Dublin Core Terms, RDA, and the FRBR model are all returned to in Chapter 3.

There are also a range of ontology languages, or meta-ontologies (Stewart, 2011, 126), 'formal languages used to construct ontologies' (Kalibatiene and Vasilecas, 2011). Each of these languages may allow for different levels of expressiveness and comprehensiveness, and there have been a number of comparisons of the different languages over the years (e.g., Gómez-Pérez and Corcho, 2002; Kalibatiene and Vasilecas, 2011). Whilst there are a number of traditional ontology languages and web-based ontology languages, and there will undoubtedly be new entrants into the market in the future, the ontology languages focused on in this book are primarily the W3C recommendations for the semantic web: Resource Description Framework (RDF), RDF Schema (RDFS), and Web Ontology Language (OWL). In Warren et al.'s (2014) survey of ontology use, of the 65 respondents answering the question of which language they used, 58 stated OWL, 56 RDF and 45 RDFS. There are well known ontologies that have been published in other languages, e.g., SUMO was written SUO-KIF, itself a variation of the Knowledge Interchange Format (KIF), (Niles and Pease, 2001) and OpenCyc makes use of Cycl (Matuszek et al., 2006), but the potential of the semantic web for bringing together distributed data means that there is often a semantic web version of the ontologies too. The structuring of the semantic web is returned to in more detail in Chapter 2.

Ontologies, metadata and linked data

The definition of an ontology provided above overlaps with both metadata and linked

data, and it is important to recognize the similarities and the differences between the different concepts, and how they overlap.

Metadata is generally defined as 'data about data', and information professionals within cultural heritage institutions have traditionally focused heavily on the creation of metadata to describe the objects within their respective collections. Extensive standards and methodologies have independently been created for cataloguing and classifying objects within each type of institution, whether archive, museum, or library, with the metadata elements reflecting those aspects considered most important within the community's culture. This may be the importance of the fonds to the archival community, reflected in the ability of Encoded Archival Descriptions (EADs) to not only describe an archive collection but also increasingly smaller parts of the collection in a hierarchical fashion, or through the extensive history of a specific object that is possible through the Categories for the Description of Works of Art (CDWA).

The traditional distinction between metadata and data breaks down, however, as we move from real-world objects to digital objects and many (often computer scientists) will say there's no point in distinguishing between the two, it's all just data. As van Hooland and Verborgh (2014, 3) put it: 'Just as you can always add an extra Lego piece on top of another, you can always add another layer of metadata to describe metadata.'

Within this work the term metadata is limited to its traditional sense, a set of elements used to describe a distinct resource, not a part of the resource itself. Where the resource that is being published is a dataset, and if the dataset has been published as linked data and the metadata has been published as linked data, then it may be meaningless to distinguish between the two.

Linked data is the best practice for publishing structured data on the web (van Hooland and Verborgh, 2014), which is generally agreed to be in accordance with the four linked data principles set out by Tim Berners-Lee:

1. Use URIs as names for things
2. Use HTTP URIs so that people can look up those names
3. When someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL)
4. Include links to other URIs so that they can discover more things

Berners-Lee, 2006

Linked data is an approach to data interoperability which offers an alternative to having an upper ontology (Murdock, Buckner and Allen, 2012). It cuts through the complexity of understanding the relationships between different terms for types of object and attributes used within different data sets by allowing the direct linking between the terms and instances rather than understanding the relationship via an

upper ontology. It is not necessary to know that 'watercolourist' and 'oil painter' are linked via the concepts 'painter' or 'artist' – instead the person J. M. W. Turner in one data set may be linked to the person J. M. W. Turner in the other directly.

It is important to recognize, however, that not all ontologies are encoded as linked data, and not all linked data is an ontology. An ontology does not have to be published on the web or necessarily follow the graph data model of the semantic web's Resource Description Frame (RDF); instead it may only be used on a private network (or even a single computer) and follow a proprietary format. Alternatively, a wide variety of data may be published as linked data without being an ontology, although when linked data may be considered an ontology and when it isn't is open to debate.

Two factors that may be used in distinguishing between linked data that is an encoded ontology and linked data that isn't an ontology are dynamism and exhaustiveness. An ontology is a formal representation of knowledge – it is not the same as a dynamic database of information; whereas the library catalogue may be considered an ontology data set or knowledge base, with rich relationships between authors and their works, the circulation aspect of an integrated library system would not be. 'Formal' also suggests that an ontology is not an ad hoc piece of data marked up as linked data; marking up the relationships between all the members of the Pre-Raphaelite Brotherhood in accordance with a particular element set might be considered an ontology, whereas someone marking up the contact details on their website would not be (although the element set used to make up the contact details might be).

What can an ontology do?

Hedden (2010, 15) identifies three principal purposes for taxonomies, each of which equally applies to ontologies: indexing support, retrieval support, and organization and navigation support. In addition to which, an ontology can also act as a knowledge base.

Indexing support

Despite advances in automatic indexing, human cataloguing and indexing continues to be an important part of the information profession, and controlled vocabularies can ensure consistency in the terms that are applied. An ontology enables an indexer to think more broadly about the terms that are applied, with a wider range of associated terms applicable.

Retrieval support

Information retrieval is the other side of indexing and cataloguing; the same terms that are used to index a document can then be used to retrieve it. The ontology, however, has a couple of advantages over other controlled vocabularies: less ambiguity and the potential of complex queries and inference. All controlled vocabularies are designed to be as unambiguous as possible, distinguishing between potentially confusing terms through the use of subdivisions, attributes, and scope notes. They are, nonetheless, subject to human error, both in their design and their implementation, and a richer set of relationships with other terms offers less room for ambiguity.

The rich set of relationships within ontologies also allows for more complex queries to be created for information retrieval. Whereas traditional search is built upon Boolean operators and faceted search, ontologies allow for increasingly complex graph matching.

Ontologies can be represented by a graph consisting of concepts and the relationships between them; for example, Figure 1.2 shows the twelve apostles of Jesus and the relationships between them as a graph.

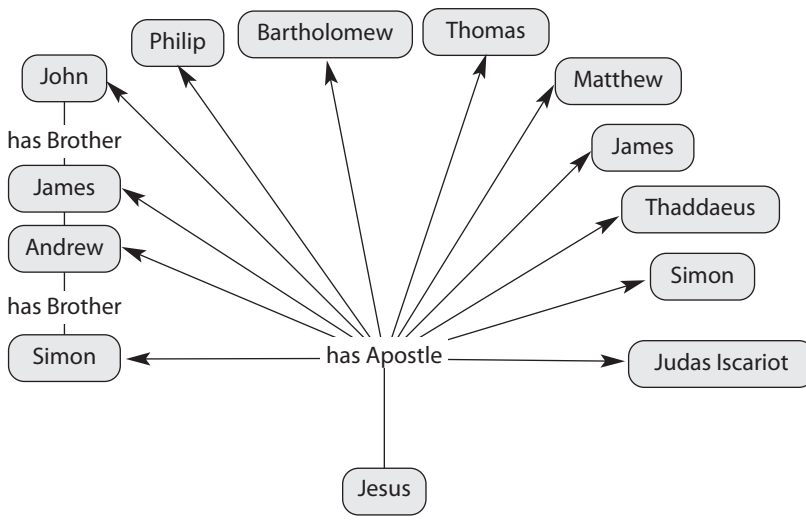


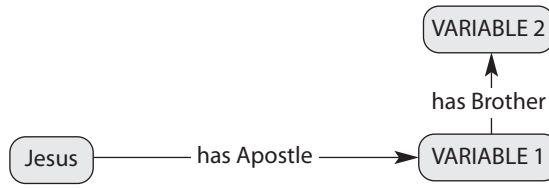
Figure 1.2 *A graph of Jesus and his twelve apostles*

For the sake of ease, within Figure 1.2 each of the people is represented by his name rather than a unique identifier which has a name as an attribute, and the fact that Simon (brother of Andrew) was subsequently called Peter is overlooked. This simple graph only includes two types of relationship 'has Apostle' and 'has Brother', and yet already graph matching enables the retrieval of results for more complex queries. If such an ontology

had been used in the cataloguing of a set of religious texts that have been variously ascribed to Jesus and his apostles it would now be possible to retrieve results as well as information from the ontology, by matching query graphs against the knowledge graph and including variables for the unknown data that the query should retrieve.

Matching the following graph against the graph of Jesus and his twelve apostles would retrieve all the apostles for the unknown VARIABLE:

Simon, Andrew, James and John would be found to match the unknown VARIABLE 1 (and VARIABLE 2) in the following graph:



The graph matching doesn't have to be built on explicit relationships alone, but may also be built on inferred relationships.

'Inference' refers to the drawing of new relationships from a data set based on existing relationships and a set of rules. At its simplest it may be an understanding of what type of thing an entity is, based on its relationship with something else. For example, if in a bibliographic ontology the information `<Charles Dickens><has written><Hard Times>` is encoded, and the relationship 'has written' is only being used to express the relationship between an author and a work, then the fact that Charles Dickens is an author and Hard Times is a work can be inferred from the information.

More extensive rules can allow for greater inference. For example, a genealogy ontology may encode two facts, that `<Adam><has Son><Cain>`, and `<Adam><has Son><Abel>`. If, as is normally the case, the `<has Son>` is only used where the target is a male, it may be inferred that both Cain and Abel are male. An additional rule stating that sons of the same parent are brothers would also allow this information to be inferred.

Organization and navigation support

'Organization and navigation support' is about the ability to find information through browsing rather than searching, following the relationships between terms to find related concepts. For example, the online store Amazon.com has an extensive

taxonomy through which a shopper may browse all the way from the general to the highly specific:

```
Books
     Politics & Social Sciences
          Social Sciences
               Library & Information Science
                    Library Management
```

Without much experience of a particular taxonomy it may be difficult to find the desired subject in an extensive taxonomy. Different people will inevitably make different decisions about the structure of a taxonomy for similar materials, and users of the taxonomy will have to learn the taxonomists' idiosyncrasies. For example on the Amazon.co.uk site 'Library & Information Sciences' is found under 'Reference' rather than 'Social Sciences' (as it is on Amazon.com) and has no further subdivisions:

```
Books
     Reference
          Library & Information Sciences
```

– whilst Amazon.ca contains 18 narrower terms than 'Library & Information Science' in comparison to Amazon.com's five. Although a taxonomy may be wrong in many different ways, there is no single correct taxonomy.

An ontology has a more complex set of relationships than a thesaurus, which creates additional challenges for enabling the browsing of resources. Whereas a thesaurus may be kept separate from the content, e.g., running down the side of the page, an ontology may be incorporated throughout the structure of the page. For example, the BBC has developed the Programmes Ontology element set (www.bbc.co.uk/ontologies/po) to facilitate access to the vast data set about the corporation's programme output and associated individuals, this information is encapsulated within a whole web page rather than one small part of it. Visiting the regular URI for the long running radio soap opera *The Archers* (www.bbc.co.uk/programmes/b008ncn6) will provide a typical HTML page of information about the series; adding .rdf to the end (www.bbc.co.uk/programmes/b006qpgr.rdf) will provide the underlying information in a machine-readable format.

The ontology as a knowledge base

An additional purpose, unique to ontologies amongst controlled vocabularies, is the

ontology as a knowledge base. The rich web of knowledge within an ontology and the ability of inferences to be drawn on existing relationships mean that ontologies can be a rich store of knowledge, not just a means to retrieve knowledge from resources indexed with a particular ontology. Certain general ontologies, such as DBpedia, draw together a wide range of information into one data set, and may be queried to produce results in a form that has not been compiled previously.

Current approaches to information retrieval are limited in their ability to discover new information (Stock et al., 2012).The use of an ontology as a knowledge base, as well as increasingly sophisticated information retrieval, is also likely to help with the discovery of undiscovered public knowledge. Undiscovered public knowledge is the idea that the discovery of new knowledge does not have to be based on the investigation of the real world of physical objects and events, but also through the interrogation of objective knowledge. Swanson (1986) identifies three forms this undiscovered public knowledge may take:

1) A hidden refutation: the hypothesis and its refutation may not both be known to any one person.
2) A missing link in the logic of discovery: if no one person knows that A causes B, and B causes C, then the inference that A causes C cannot be known.
3) Combination of multiple tests: a meta-analysis of multiple weak tests may nonetheless provide a strong result.

Each of these is fundamentally an information retrieval problem: ensuring that both hypothesis and refutation are found by a search; ensuring subsequent statements are found; ensuring that all available tests of sufficient quality are identified. Ontologies can undoubtedly improve information and knowledge retrieval, and help with the mining of undiscovered public knowledge, in an increasingly automated fashion.

Ontologies and information professionals

This book is being published at a pivotal point in the history of ontologies. On the one hand the web and the development of semantic web technologies have provided the opportunity for ontologies to be adopted by more people in more places than ever before – bringing together data from around the world into one huge data set that can be queried by anyone. On the other hand the ideals of a semantic web have had to adapt to the practicalities of human abilities, recognizing the importance of publishing data even if it is not accompanied by robust formal ontologies.

This book will not only emphasize the importance and potential of ontologies, but also the importance of the community of information professionals contributing to

the development of new, and increasingly useful, ontologies. Murdock, Buckner and Allen (2012) point out that one of the problems with ontology development is the need for 'double experts', those with knowledge of ontology design and subject domains. The community of information professionals have a long tradition of being 'double experts', often coupling a postgraduate information science degree with a subject specialism, and are ideally placed for a role in facilitating access to the web of data and the development of ontologies. The role is particularly important if we are to avoid the risk that an ontologist's imposition of a domain ontology masks how practitioners construct meaning (Pike and Gahegan, 2007).

Knowledge and experience of using knowledge organization systems is a prerequisite for many jobs within the information profession, and the need for knowledge of ontologies more specifically, is only likely to increase in the future. As well as taxonomists and ontologists, for whom the development and maintenance of controlled vocabularies may be a full-time role, knowledge and experience of ontologies is also necessary as part of a wider skill set in cataloguing, metadata and curation roles. For those working as a taxonomist for a global information service, a metadata librarian in a university library, a digital asset cataloguer in a commercial company or a records manager in a non-profit organization, it is increasingly difficult to overlook the importance of ontologies.

The focus of the ontologies in this book is on those that are being used on the semantic web. There are, of course, many bespoke and proprietary ontologies used within commercial organizations, attempting to bring together the disparate information created by departments and units, but those that are of greatest interest are those that provide the opportunity to share more data than ever before and develop new insights from across the world.

Alternatives to ontologies

It is important to recognize that ontologies have limitations, and that there are alternative ways of capturing and analysing data. Some of the limitations can be traced to the fundamental assumptions that are made when encoding knowledge within ontologies. Brewster and O'Hara (2007) note two such assumptions: first, the monolithic nature of knowledge that is continually added to; and second, that concepts are the fundamental units of ontologies, and these are manipulated with language.

Although there may be few, if any, Kuhnian paradigm shifts (Kuhn, 1970) that invalidate the whole of an ontology, there will nonetheless be changing perspectives on the meanings and relationships of individual concepts. This is especially true outside the sciences, where the meaning of concepts and the relationships with associated concepts can be open to vigorous debate. There is also much that is difficult

or impossible to put adequately into words –so-called tacit knowledge (Polanyi, 1966) – although Shadbolt and Smart (2015) suggest that rather than tacit knowledge being seen as something that is impossible to articulate, it should be seen as something that is more easily articulated in some situations than others.

Approaches to knowledge representation can be broadly categorized as either top-down or bottom-up (Pike and Gahegan, 2007). Whereas ontologies can often be considered top-down models of the world, especially when considering the creation of universal ontologies such as OpenCyc, the development of linked data and the semantic web allow for a more bottom-up approach with competing ontologies and potentially conflicting perspectives. However, even bottom-up approaches to capturing knowledge from the data that is available have limitations.

The sheer quantity of information available on the web provides, and necessitates, alternative ways of capturing data, through automatic reading and natural language processing (NLP). NLP can be used both to extract terms for an ontology or thesaurus, and apply terms from an ontology or thesaurus during indexing; the difference between structured and unstructured data is becoming increasingly blurred (van Hooland and Verborgh, 2014). NLP has its limitations, however, and depending on the content and purpose of the NLP it is better categorized as a semi-automatic rather than an automatic process. NLP is not the principal subject of this book, but it is likely to play an increasing role in the development of ontologies in the future, and the subject is returned to in Chapter 5.

Neither the limitations of ontologies, nor the alternatives, dismiss the need or importance of ontologies. Rather, they help us understand where and when ontologies are appropriate. It may be that in some situations a simpler form of controlled vocabulary is more appropriate, either a thesaurus or an authority list. Data may be better stored in a list, a spreadsheet or a relational database than as a graph, whilst certain types of tacit knowledge may be better captured through video than by trying to put it into words. Brewster and O'Hara (2007) note that criticisms have been made that ontologies demand too much work and are too rigid, but such criticisms have been made about many core information activities, such as cataloguing and classification in the age of the web, and what we find is that most often new technologies complement rather than replace existing technologies. Rather than search engines replacing the library catalogue, the library catalogue is increasingly integrating its own information services with the web and, increasingly, the semantic web. Rather than ontologies replacing earlier forms of controlled vocabularies, they complement them, providing an increasingly powerful tool for information retrieval and knowledge representation.

The aims of this book

There are three main aims for this book. The first is to demonstrate to the information professional the importance of ontologies for knowledge discovery. The second is to demonstrate the important contribution information professionals can make to the development of ontologies. Finally, the book aims to provide a practical introduction to the development of ontologies for information professionals.

This introductory chapter will, hopefully, already have gone some way to demonstrating the importance of the development of robust and widely used ontologies in the fight against information overload, and the role of the information professional in the process. These ideas will continue to be developed and reinforced throughout the rest of the book.

In addition to demonstrating the importance of ontologies and the role of the information professional, the book is also designed to be a practical introduction. It will introduce some of the existing dominant ontologies that are likely to be of interest to the information profession, as well as the methods and tools necessary for building new ontologies and interrogating existing ontologies. LaPolla's (2013) survey found the implementation of semantic web compliant catalogues was hindered by a lack of funding, best practice and awareness of the associated concepts. Whilst the book can do little about the lack of funding, it will contribute to both discussion on best practice and increase familiarity with many of the basic concepts. Although a majority responding to LaPolla's survey had some familiarity with semantic web concepts, this fact is clouded by the fact that it was a self-selecting survey and it seems likely that those with little interest in the semantic web didn't bother with the survey. Even amongst those who completed the survey, whereas the vast majority were either very familiar or somewhat familiar with the concept of the semantic web (90.16%) and linked data (95.52%), familiarity with more specific technologies necessary for implementation were far lower: Web Ontology Language (OWL), 53.21%; Simple Knowledge Organization Systems (SKOS), 43.59%.

No single book could provide an exhaustive introduction to the practicalities of ontology use and development. Whole books have been written on technologies that have been covered here in one or two pages; there is a huge variety of software available for ontology development; new ontologies are being developed (as well as old ones falling into disuse); and old standards are changing while new ones are introduced. Nonetheless, the underlying methods of ontology development change more slowly than the specifications, and by focusing on the underlying theory the skills related to one set of technologies can be applied to others.

The structure of this book

The rest of this book consists of six chapters, from introducing the semantic web and some existing ontologies, through adopting, building and interrogating ontologies, to the future of ontologies:

Chapter 2 – Ontologies and the semantic web

Ontologies have gained added significance in recent years through the adoption of an increasingly semantic web. Chapter 2 provides an introduction to the semantic web and the role of ontologies, and how ontologies have been increasingly adopted in a wide variety of libraries as well as other cultural heritage institutions and commercial organizations.

Chapter 3 – Existing ontologies

There is a wide variety of ontologies that have been developed, and knowledge of the dominant ontologies, their applications and their differences is increasingly essential to the information professional. Chapter 3 considers some of the main ontologies, including those ontologies used for representing ontologies, those widely adopted by libraries and those widely used on the web.

Chapter 4 – Adopting ontologies

The reuse of existing ontologies is important for both the integration of data across different systems and to avoid the repetition of work. Chapter 4 considers the tools that are available for identifying existing ontologies, how the ontologies (or elements thereof) can be combined in the creation of application profiles, and some of the criteria that should be considered when selecting ontologies.

Chapter 5 – Building ontologies

It is increasingly important that information professionals are not only users of existing ontologies, but that they build their own ontology for particular applications. Chapter 5 provides both a methodology for building an ontology and an overview of some of the tools that are available, before leading the reader through the development of a simple ontology with Protégé, the most popular (and free) software for ontology development.

Chapter 6 – Interrogating ontologies

Ontologies are not only of interest for the structure they provide, but also for the data that they contain. Chapter 6 provides an overview of tools available for interrogating semantic web ontologies, both through Simple Protocol and RDF Query Language (SPARQL) and web crawlers, to gain new insights.

Chapter 7 – The future of ontologies and the information professional

The final chapter looks to the future of ontologies and the role of the information professional in their development and use. The future of ontologies will undoubtedly be a mixture of lightweight and more formal ontologies, and their development is likely to be integrated with other technologies such as Natural Language Processing and potentially crowdsourcing workflows. The contribution for the library and information professional to ontology development also has the potential to change, expanding from the bibliographic ontologies that will undoubtedly occupy them in the short term to the development of niche subject specific ontologies in the long term.