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.

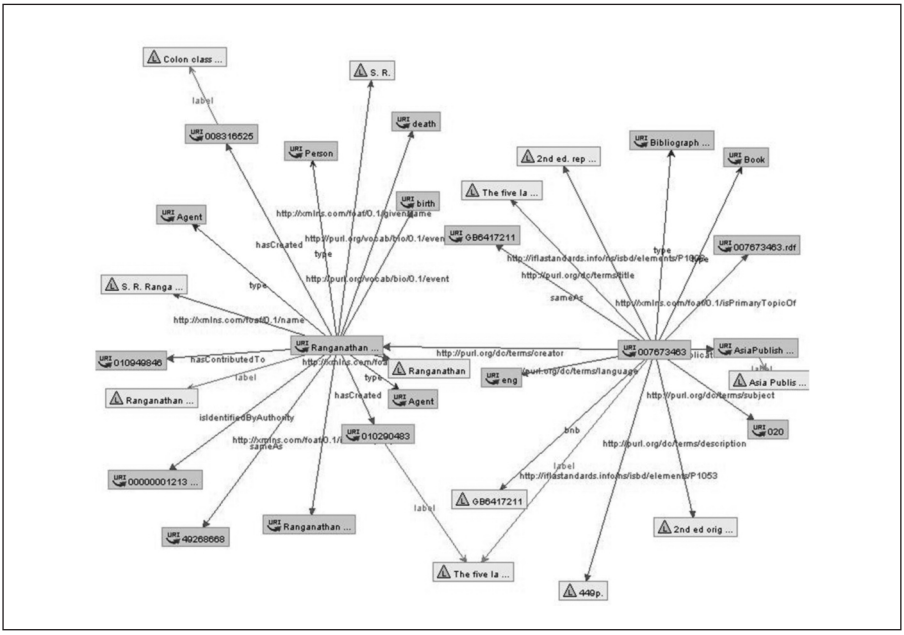

Figure 1.1 *Section of the British National Bibliography graph visualized using RDF Gravity*

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