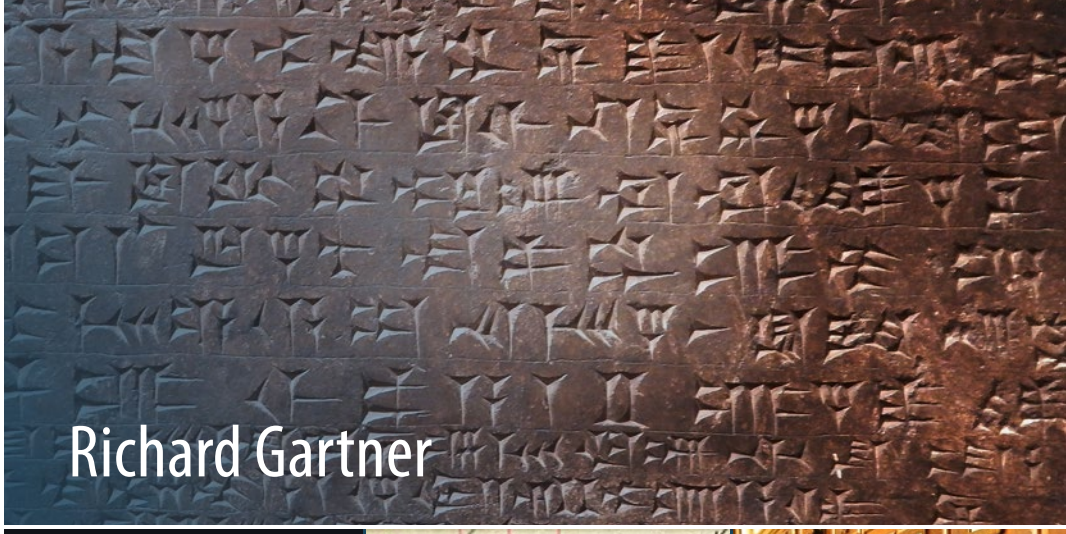
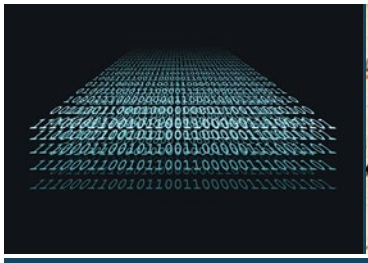
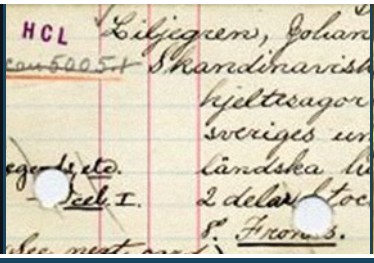


Richard Gartner





Metadata

Shaping Knowledge from Antiquity to the Semantic Web

Springer

Metadata

Richard Gartner

Metadata

Shaping Knowledge from Antiquity to the Semantic Web

Springer

Richard Gartner
The Warburg Institute
University of London
London, UK

ISBN 978-3-319-40891-0 ISBN 978-3-319-40893-4 (eBook)
DOI 10.1007/978-3-319-40893-4

Library of Congress Control Number: 2016947721

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG Switzerland

Acknowledgements

Many friends and colleagues past and present have helped me build up my understanding of metadata over the years – too many to list separately. But I would like to give special thanks to those who have provided such helpful feedback on my initial proposal and an early draft of this book: Sheila Anderson, Sarah Bakewell, Tobias Blanke, Jonathan Blaney, Anna Jordanous and Gareth Knight – thank you.

I would also thank those who have so generously given their permission to reproduce their works in this volume: Michael Bergman, John Blyberg, the British Library, CERN, King's College London Library, the National Film Board of Canada, the National Library of Australia, Jenn Riley, Tim Watts and the Zooniverse project.

Thanks are also due to those who have generously made their works available under Creative Commons or other open licences to allow their reproduction here: Markus Angermeier, Christian Bizer, Luca Cremonini, Richard Cyganiak, Mark Doliner, the Education Resources Information Center, Anja Jentzsch, Montréalais, Aaron Rotenberg, Max Schmachtenberg, Daniel R. Strebe, the Wikimedia Foundation, George H. Williams and Richard H. Zander.

I would also like to thank Beverley Ford, James Robinson and S. Madhuriba at Springer for guiding me from initial proposal to publication with such smooth efficiency.

And finally more than mere thanks are due to Elisabetta Flamini and Ann Grupman who have always given me help and support that is so much more than 'meta-'.

Contents

Chapter 1
What Metadata Is and Why It Matters

In May 2013 headlines were made around the world when Edward Snowden, a former employee of the United States Central Intelligence Agency revealed that the country's National Security Agency (NSA) had been carrying out surveillance on the email and mobile phone activities of millions of people at home and abroad. The net of those under scrutiny was cast so far and wide that it had even picked up some world leaders in its catch: they included Germany's Angela Merkel who was so incensed that she publicly compared the NSA to the Stasi, the secret police force of the former East Germany [1].

What rapidly emerged from the flurry of headlines that came from these revelations was that the NSA was not employing such classic techniques as phone tapping or the interception of emails, but was instead collecting information *about* phone conversations, text messages or emails. The news media began using the term *metadata* to describe information of this type: this was not what was actually said in a phone call or written in a text message but such details as the location of the phone used, the numbers called, the time and dates of calls and so on. Because it was metadata that was being collected, critics of Snowden's actions claimed that the NSA's actions were not overly intrusive and could be accepted as the price of national security [2].

Metadata became a commonly-used term in the media during the Snowden affair. But what exactly is metadata? *The Guardian* newspaper attempted to clarify it for their readers when they published *A Guardian guide to your metadata* in June 2013: metadata, they said, is "information generated as you use technology…[not] personal or content-specific details, but rather *transactional* information about the user, the device and activities taking place" [3] [italics added]. The information they referred to in the article is the type generated by Twitter when a Tweet is posted. Figure 1.1 shows something of what it looks like.

Even this small sample is enough to show that metadata contains more than *The Guardian's* simple definition allows for: it moves well beyond information on the 'transaction' of posting a Tweet and enters the realm of the personal. From it we can tell where its author is located, how many followers, friends and favourites they

R. Gartner, *Metadata*, DOI 10.1007/978-3-319-40893-4_1

```
  'created_at' => 'Mon Jun 22 10:45:09 +0000 2015',
  'id' => '612934343330127873',
  'user' => {
      'id' => 2123456,
      'name' => 'AnExample',
      'screen_name' => 'AnExample',
      'location' => '"London"',
      'followers_count' => 891,
      'friends_count' => 516,
      'listed_count' => 61,
       'created_at' => 'Wed Feb 04 07:50:12 +0000
2009',
      'favourites_count' => 143,
      'time_zone' => 'London',
      'description' => "Frequent contributor to
Twitter"
   }
```

Fig. 1.1 Metadata from a Twitter feed (fictitious and heavily truncated example)

have and when they opened their account; we can also read a description of themselves that they added to their Twitter profile. Obviously there's more to metadata than 'transactional' information alone.

Etymology is always a good place to start defining a concept and metadata is no exception. The prefix *meta-* comes from ancient Greek and is usually translated into English by the preposition *about.* It is often used to express an idea that is in some way self-reflexive. In linguistics, for instance, a *meta-language* is a language that describes another language. In mathematics, a *meta-theory* is a theory about a second theory. In literature a *meta-fiction* is a fictional work, such as Laurence Sterne's *Tristram Shandy,* which draws the reader's attention to its status as fiction. So, unsurprisingly, *metadata* is usually defined as *data about data*.

What is usually accepted as the first use of the term in computer science occurs in a research publication by the US Air Force from 1968. It occurs in a technical paper called *Extension of Programming Language Concepts* by the computer scientist Philip Bagley. In it, he attempted to define models for new programming languages that were not intimately tied to particular types of computer hardware (as had previously been the norm). An essential component of a programming language, he claims, is:-

> the ability to associate explicitly with a data element a second data element which represents data "about" the first data element. This second data element we might term a "metadata element" [4].

The examples Bagley gives for this new concept include identifiers or labels for pieces of data, 'prescriptors' which limit the range of values a data element can have and codes which limit how and when it can be accessed. All of these are clearly *data about data* and would be recognized as metadata today.

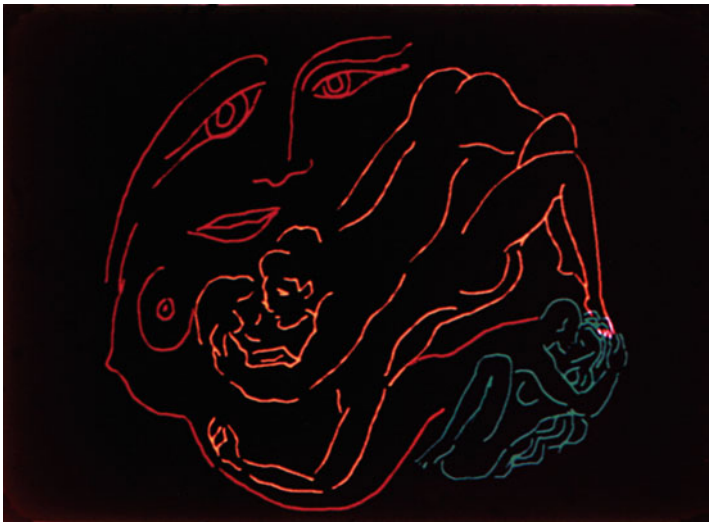


Fig. 1.2 Metadata ©1971 National Film Board of Canada. All rights reserved

The term was rapidly adopted by practitioners of computer science and appears frequently in research papers and publications from the 1970s onwards. It even became well enough recognized outside this community to have become the title of an animated film made by the National Film Board of Canada in 1971 (Fig. 1.2).

The title was no doubt partly chosen because of the film's pioneering use of computer animation to render its flowing line drawings. But perhaps it also alludes to the abstraction of its images, their reduction of reality to simple line components. The animator Peter Foldès' paring down of the real world in this way reveals a key feature of most metadata, that it is selective and throws away much of what it could potentially say about the data it describes. What is chosen and what is ignored is a key part of defining metadata and how it is used.

Although the term itself has a provenance of only half a century, metadata has clearly been around much longer and has had a life well before the computer was invented. One group who have been producing it in large quantities for millennia are librarians: they have been listing and cataloguing their collections for much of recorded history. The concept of metadata adapts easily to a metaphor for the way in which a library works: its 'data' is formed from the collections housed within it, its 'metadata' from the catalogues used to describe them. Because the creation of metadata has always been a core part of a librarian's job, it is not surprising to find that many of the innovations in its history have their origins in libraries.

A Human Construct

One feature of metadata that is often forgotten is that it is a human construct and not found in nature. The shape of metadata is designed by human beings for a particular purpose or to solve a particular problem, and the form it takes is indelibly stamped with its origins. There is nothing objective about metadata: it always makes a statement about the world, and this statement is subjective is what it includes, what it omits, where it draws its boundaries and in the terms it uses to describe it.

To illustrate this, we can look at a simple metadata representation of our own Earth. We could think of our planet as an immense body of data: its physical features are too complex and on too large a scale for us to grasp them from our observations of the small part we experience every day. To make sense of the world, we can use an object such as a globe (Fig. 1.3) which abstracts its most important features to make them more intelligible to us. Although some parts of a globe mirror what is found in nature (the shape of land masses for one), many of the details on its surface add human-created metadata to these representations of physical reality. The real world has no lines of longitude or latitude drawn on it and no physical counterpart for its political boundaries or the colour-coding of the entities they delineate; it also has no English-language labels covering its surface.

All of these are added for a purpose: the lines of longitude and latitude provide a useful grid for locating points on the Earth's surface, and the political boundaries reflect a human-made division of the planet that has evolved over lengthy periods of time. The designer of this artefact has been selective about what to include and what

Fig. 1.3 A globe: metadata as a human construct (Photograph by Mark Doliner)

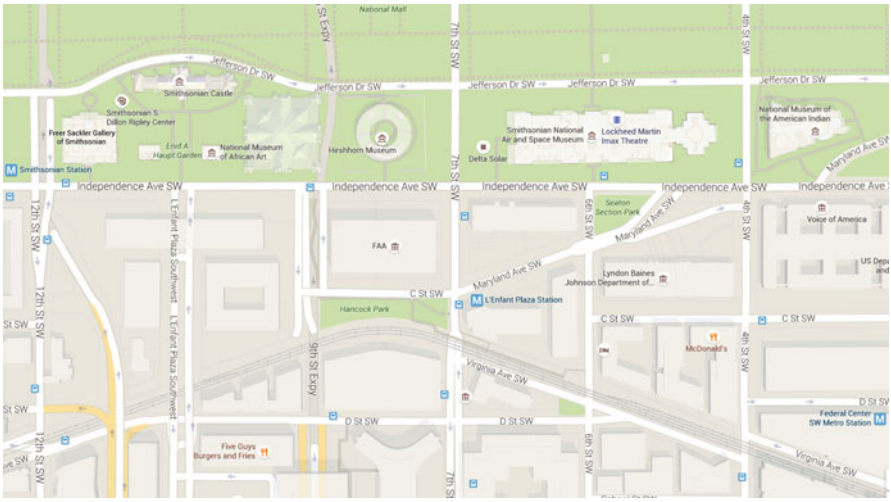


Fig. 1.4 Map of central Washington DC (Google maps)

to omit: equally valid representations may, for instance, leave out political boundaries and concentrate on the physical or environmental features of the planet. The terms used to describe the features shown are also selective: here they are labelled mainly in English or in forms easily comprehensible to English speakers. This globe clearly shows the purpose for which it was designed and the (English-speaking) community within which it originated. All metadata bears the imprint of its origins and intended use when examined closely enough.

The purpose for which metadata is constructed can drastically alter the form it takes even if it is describing the same data. A map of central Washington D.C. produced by Google maps (Fig. 1.4) appears to offer us an objective picture of the most significant features of this urban landscape. It shows the most important buildings, the roads that connect them and such useful landmarks as the location of Metro stations. This appears to be an authentic and accurate summary of this space, but it remains a highly selective one: only some features of the landscape are present and their number decreases the further one zooms out.

The ones which are shown are there because they help the map meet the purposes for which it is designed. Its primary function appears to be navigation: it gives priority to the location of sites on the ground and to the roads by which we can walk or drive to them. The overall shape of the map and the ways in which it represents the spatial distribution of the features it describes follow long-established cartographic conventions such as putting north at its top edge.

By comparison with Google Maps, a map of the Washington Metro (Fig. 1.5) offers a very different view of the same portion of our planet. Here the amount of metadata included is drastically pared down to little more than the names of stations and the lines that connect them. The pattern of the streets above ground is ignored and the spatial geography of the stations, correctly shown on the Google map, is

Fig. 1.5 Central section of Washington Metro rapid transit system map (truncated) (Original diagram by Montréalais)

rendered completely inaccurate. But this map works very well as metadata for its intended purpose: it is designed to provide a simple way to navigate the public transport network of Washington and so gives priority to expressing the information required for this and nothing more. A passenger does not need to know the exact geography of the stations above ground, just the lines needed to reach them and those at which connections can be made: everything else can be omitted and the map still fulfil its purpose.

Three Types of Metadata

Metadata can serve a complex assortment of purposes but most of these divide into three clear categories: this tri-partite division can be useful to help us make sense of what a particular type of metadata is trying to do. Each of these is worth looking at in some detail.

The first, usually called *descriptive* metadata, is the one with which we are perhaps most familiar as it tends to be the most conspicuous. This is the metadata designed to help us discover and locate the data it refers to. In a library, it takes the form of the information we see when we look at a record in a library catalogue: it may include such details as the author and title of a book, who published it, when it was published and what its subject is. It may also cover information that can help us to identify it unambiguously, such as its ISBN (International Standard Book Number).

Because its main purpose is to allow us to locate an item, descriptive metadata is often referred to as *finding metadata*. It should let us find something whether we know exactly what we're looking for, for instance if we're after a particular book for which we've read a review, or whether we have little more than a rough idea of the subject in which we are interested. It should do this even if we are just browsing serendipitously without a particular work or topic in mind. It also, and just as importantly, allows us to *exclude* something from our search by providing information that allows us to assess its relevance. In a library catalogue, this may be as simple as keywords describing the subject of the book: in more commercial systems, such as the online seller Amazon.com, it can be much more elaborate, often including reviews by other users and recommendations generated by the system on the basis of our searching habits.

Descriptive metadata may be the variety with which we are most familiar but it is often dwarfed in size and scope by another type which is needed to keep any information system (electronic or otherwise) running. This is usually called *administrative metadata*, the background information that ensures data can be stored, preserved and accessed when it is needed. Metadata of this type has always been necessary: every library has had to keep records to allow its administration to operate. In the electronic environment this metadata is larger in volume and often much more complex in content.

Some of this is the technical information needed to allow a system to operate and serve digital data up to us in a usable form. This type, unsurprisingly called *technical metadata,* covers everything a system needs to know about a digital object to deliver and render it properly. For a digital image this may include details of its size in pixels, the file format to which it conforms, the size of the palette of potential colours each pixel may use, any compression used to reduce its size and many more bits and pieces. Different technical metadata is needed for different types of digital object: what is required for a still image is noticeably different from that needed for video, audio or text. This metadata is usually generated automatically as a system is running and is generally invisible to the user (unless something goes horribly wrong).

Another important type of administrative metadata is the information needed to allow a system to enforce intellectual property rights (IPR). This *rights metadata* can include such elements as details of those who own the IPR, the copyright laws under which their ownership rights apply and what rights they grant to those who access their data. IPR applies, of course, outside the digital world and rights metadata is attached to almost every printed book as a copyright declaration behind the title page. In an electronic environment it is usually more complicated than simply a textual description of ownership: it is the way in which different types of access to a digital object are controlled. An online newspaper such as *The New York Times,* which charges users to read more than ten articles a month, relies on this metadata to enforce its payment mechanisms.

The final type of administrative metadata is the information necessary to make sure that our digital data will be accessible and usable well into the future. This *preservation metadata* is intricate and extensive because the processes for preserving

digital information over long periods of time are themselves highly complex. The hardware and software in use today have a very short shelf-life in comparison to the potential timespans over which our data may remain valuable. Our digital heritage needs careful maintenance to remain viable over these long periods. Supporting the processes to ensure this requires large amounts of metadata that documents what is done to the data, how it is done and by whom, all of which ensures it can be used when its original creators (machine, software and human) are long gone. As with most administrative metadata, this is invisible to the user, but vital if a digital sinkhole is not to swallow up much of the world's data.

A final type of metadata is also not always obvious to the user but is often indispensable. This is the information that builds links between small pieces of data to assemble them into a more complex object, digital or otherwise. In a Kindle e-book, this is what tells the device to deliver page 1 before page 2 before page 3 and so on; it can also be used to group these pages into chapters. This information is called *structural metadata* as it defines structures which bring together simple components into something larger that has meaning to a user. It is what turns an otherwise random assortment of pages into something we can recognize as a book. Information of this type is inherent even in a physical book in the page numbers and the order in which the leaves are glued into the spine. In the digital world, it has to be recorded more explicitly if an object of any complexity is going to make sense as something more than an incoherent pile of data.

These three types of metadata, descriptive, administrative and structural, show that a simple definition that sees it as 'transactional information' alone covers only a small part of what it can do. Metadata is certainly a complicated business. What is also indisputable about it is how fundamental it is to humans and the ways in which their knowledge and culture is built up.

Why Metadata Is Needed

Metadata exists for a reason and that reason lies fundamentally in the limitations of the human brain. Amazing though it is in many ways, the brain is restricted in the amount of information it can store and retrieve accurately. There are ways around this, of course. One trick that many learn is to use simple mnemonics to commit to memory abstract concepts that are otherwise difficult to retain: the colours of the spectrum are easier to recall if we memorize the phrase "Richard Of York Gave Battle In Vain" than if we try remember the unadorned list "red-orange-yellow-green-blue-indigo-violet". Mnemonics such as this work by pulling out key features of the data they are designed to preserve in our memory and redrawing them in a more cognitively-manageable form. They employ a key principle of metadata, abstracting patterns from the data it is 'about' to make its retrieval easier.

Once we move beyond the confines of a single brain, the need for metadata rapidly becomes ever more pressing. One of the most significant upheavals in the ways in which we preserve our memories outside the confines of the human cranium was

the invention of writing. The anthropologist Jack Goody wrote perceptively as far back as 1963 of the impact of this development, noting that for the first time it allowed culture to be transmitted relatively intact between generations. Writing, he says, obviates the ways in which orally-transmitted culture, passed through a chain of interlocked conversations, gets automatically changed because of the pressures and imperatives of social life in the present [5]. In an orally-transmitted culture, he points out, the "individual has little perception of the past in terms of the present" [5]: writing brings an awareness of "the past as different from the present; and of the inherent inconsistencies in the picture of life as it [is] inherited by the individual from the cultural tradition in its recorded form" [5].

Profound as the impact of writing was, its effects would have been severely circumscribed without some way of organizing the written record above the level of the single text. It is not enough just to write something down, it has to be found again. Once our written record grows to anything beyond a few snippets, we need to put it into some logical shape if we are ever going to do anything with it. Filing systems in some form soon become a necessity and to make them work we need metadata; at the very least we need to attach labels to what we have written down, shorter than the texts themselves, so that these can be put into some coherent order. Metadata, already a useful aid to retrieval when information is held within the human brain, soon becomes essential when it leaves its confines.

But metadata has a role that is more important than simply allowing a single fragment of information to be stored away and found again. It enables these to be linked together to form knowledge and for this knowledge to be consolidated into what we understand as culture.

From Information to Knowledge to Wisdom

Knowledge as a concept can be hard to pin down: if it weren't, there would not be a whole branch of philosophy, known as epistemology, which attempts to define it and examine the forms it takes. There has been plenty of disagreement about what knowledge is for millenia. It certainly goes back as far as Plato, who famously defined knowledge as *justified true belief* [6], a hugely influential tri-partite definition with which much epistemology has been in dialogue ever since. In the latter part of the twentieth century, other approaches to defining knowledge emerged which see it in terms of its relationship to information. This strand, begun by the American philosopher Fred Dretske in his best-known work *Knowledge and the Flow of Information* [7], still emphasizes that knowledge is in essence a type of belief, but one that is specifically grounded in information.

The notion that knowledge stems from information has naturally proved influential in the area of information science where new models have emerged that attempt to move away from associating it inextricably with belief. Key to these are attempts to gain some clarity on the relationship between data, information, knowledge, understanding and possibly even wisdom. One of the most influential is known as

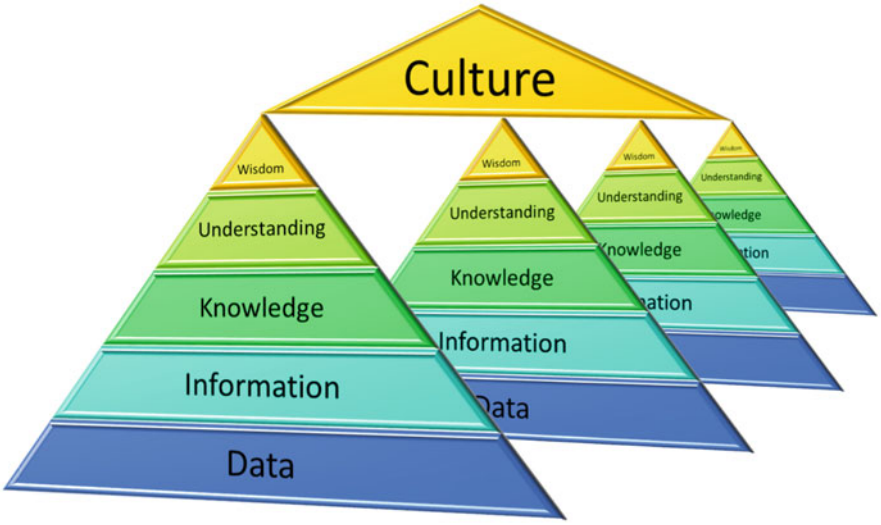


Fig. 1.6 Ackoff's Pyramid topped by a layer for culture

the *Ackoff Pyramid*, after the organizational theorist Russell Ackoff; it was he who, in an article grandiosely titled *From Data to Wisdom,* defined a pyramid of these interlocking concepts stacked on top of each other (Fig. 1.6) [8].

Ackoff's pyramid starts at the bottom with its most diminutive component, a single unit of data. This has very little meaning in itself. A single cell in a spreadsheet, for instance, is just a number, inconsequential unless we know something of its context. This can be gleaned by looking at its relationships to other components in the same sheet, such as a label at the top of the column in which it is located. Once some meaning is inferred by looking at the relations between these units, we begin to move up from data to information, the next level in the pyramid. Information may be thought of as organized data, arranged so that it can answer basic questions about its world.

The next level up from here, to knowledge, occurs when information is collected together and meaningful links are made between its components; the whole then becomes more than the simple sum of its constituents. For this to happen, patterns have to emerge which are themselves meaningful; the shape of these patterns and their relationships form the foundations of knowledge. Another way of depicting knowledge is to see it as *information in context*: a unit of information is no longer an isolated statement about the world but gains new meaning by interacting with its peers.

Interpreting and analyzing these patterns so that new ones can emerge is what allows us to move up another level, from knowledge to understanding. Much academic research attempts to do just this, establishing understanding by looking for patterns in knowledge and drawing inferences from these. The route to the next move, from understanding to wisdom, is rather harder to pin down: this is where notions of right and wrong or of the best possible alternative come into play.

Unsurprisingly, while information science generally feels comfortable with the first four levels of the pyramid, attempting to define, let alone model, wisdom is something it usually fights shy of.

Another ways of looking at this pyramid is to consider each level as offering potential answers to a different type of question. Information allows us to answer such pithy questions as "who?", "what?", "when?" or "where?". Knowledge allows us to answer "how?", particularly "how things (concrete or abstract) work?". Understanding allows us to ask "why?", particularly "why are things the way they are?" Wisdom should allow us to ask "what is best?" or "what is the right thing?". Data alone, at the bottom of the pyramid, cannot answer any of these questions by itself: it is only when it aggregated into the higher levels that this becomes possible.

In this model of information, knowledge, understanding and (possibly) wisdom, metadata has a central role in allowing Ackoff's edifice to be constructed from the atom-like units of data at the bottom of the pyramid. In fact, to move up between each of these levels relies crucially on establishing linkages between the components on one to create aggregations that become the units of the next. At every stage, metadata has a vital role to play as the 'glue' with which these links are joined.

At the bottom of the pyramid, turning data into information requires placing it into some relationship with other data. This 'relationship' is a form of metadata by any other name: it is saying something about both units of data, even if it's not clear what this 'something' is. In our spreadsheet which turns data into information, the location of a cell and its place in relation to others in the same sheet tells us something about that data and so is unmistakably 'data about data'. There is an obvious analogy here with linguistics: the 'structuralist' theories of Ferdinand de Saussure [9] argue that a sign acquires its meaning by its place in a wider structure. Metadata here is the structure which gives our signs, the units of data, their basic meaning.

In moving from information to knowledge, the role of metadata is even more explicit. As we have seen, this stage is about putting information into context and forging links between its components to generate new meanings. All metadata can do this, but it is particularly true of its descriptive form. Attaching this metadata to a component of information immediately establishes a connection to others of its kind. Labelling a book with its author's name, for instance, forges a link to other books by the same author. These are *semantic* links: they express some meaning about the relationships between the information they join together. A network of these rapidly allows us to begin to answer the 'how' questions that are the domain of knowledge.

The move from knowledge to understanding is another layer of abstraction from this network of metadata: as we have seen, understanding is a process of analyzing its patterns to derive higher-level ones that answer new types of questions such as 'why?'. We could, if we were particularly pedantic, label this meta-metadata, as it is essentially saying something about the metadata at the next level down. Similarly, when we move to the more tendentious level of 'wisdom', we are using new aggregations of metadata, aimed at answering questions of right and wrong, on top of this level of understanding.

From Knowledge to Culture: The Role of Curation

Metadata clearly has a crucial role in allowing us to aggregate data into knowledge and understanding (and maybe even wisdom). But an even higher level could be said to exist above those of Ackoff's pyramid, that of culture itself. Here, again, metadata has a role to play.

A culture is always evolving but this evolution relies on the preservation and transmission of its earlier manifestations: no culture appears spontaneously without reference to these, even if it is a rebellion against the past. To enable this to happen, a culture has to be *curated*, literally 'cared-for' as the word's etymology (from the Latin *curare*) indicates.

Curation is often confused with preservation, but there is much more to it than this alone. Curation involves identifying those elements of a culture that particularly define it and choosing which ones are important; it then describes and adds context to these, making connections between them, so that they can be understood by all who have an interest in them. Finally, it involves disseminating a culture, making it accessible. All of these are in addition to ensuring that these elements will continue to exist for a long time in the future. Going through these steps ensures above all that a culture can be understood when it is transmitted between generations. It is thanks to the curatorial efforts of our forebears that any culture beyond the most ephemeral has any existence at all.

At the core of curation lie the processes of organising and describing: whether we are curating a priceless artefact in a museum or a Twitter feed in a digital archive, these two processes are key to putting culture into context. To do this requires metadata, the same as is needed to condense information into knowledge but here with a specific emphasis on classifying, establishing linkages and explaining. This new role for metadata can be seen as an extra layer on top of Ackoff's pyramid, joining multiple aggregations of knowledge, understanding and possibly wisdom into a wider culture.

The role of metadata in curation will become clearer as we look into its history, but it remains as important as ever today in ensuring our culture's continued existence and evolution. This is particularly so when so much is digital and liable to disappear into an electronic void unless special care is taken to prevent this happening. The extent of our digital culture and the speed with which is changes are both of a different order of magnitude from anything we have experienced before but the need for metadata to ensure it remains curated is as acute as ever.

Because metadata has been so fundamental in helping us preserve human culture beyond the ephemeral moment, it makes sense to start looking at its history from long before information and knowledge became digital. To begin the story, it is time to meet one of the world's earliest curators and creators of metadata, Princess Ennigaldi of ancient Babylon.

References

1. Traynor, I., & Lewis, P. (2013). Merkel compared NSA to STASI in heated encounter with Obama. *The Guardian*. http://www.theguardian.com/world/2013/dec/17/merkel-compares-nsa-stasi-obama. Accessed 3 Sept 2016.
2. British Broadcasting Corporation. (2013, July 6). Transparency and secrets. *Moral Maze.*
3. Guardian, U. (2013). A Guardian guide to your metadata. *The Guardian*. http://www.theguardian.com/technology/interactive/2013/jun/12/what-is-metadata-nsa-surveillance. Accessed 21 Jan 2016.
4. Bagley, P. R. (1968). *Extension of programming language concepts*. Philadelphia: University City Science Center.
5. Goody, J., & Watt, I. (1963). The consequences of literacy. *Comparative Studies in Society and History, 5*, 304–345.
6. Fine, G. (2003). *Plato on knowledge and forms: Selected essays*. Oxford: Oxford University Press.
7. Dretske, F. (1981). *Knowledge and the flow of information*. Cambridge, MA: MIT Press.
8. Ackoff, R. L. (1989). From data to wisdom. *Journal of Applied Systems Analysis, 16*, 3–9.
9. de Saussure, F. (1993). *Course in general linguistics*. London: Duckworth.

Chapter 2
Clay, Goats and Trees: Metadata Before the Byte

Ancient Metadata

In the course of his excavations in the ancient city of Ur during the 1920s the archaeologist Leonard Woolley made a discovery which initially puzzled him: in the remains of a chamber of the sixth-century BCE palace complex were scattered an array of much older objects. Some of these pre-dated the room in which they were found by 700 years, others were 900 years older than that; all were neatly arranged next to each other in what seemed a careful ordering. Woolley soon concluded that he has stumbled on a museum, one created and curated by Ennigaldi-Nanna, the daughter of the last king of the Neo-Babylonian empire.

The convincing evidence for Woolley was the discovery of a small number of clay cylinders amongst the remains. Each of these was a 'small drum-shaped clay object on which were four columns of writing; the first three columns were in the Sumerian language…the fourth column was in the late Semitic script' [1]. He had found, he thought, the first example of a museum label, probably created by the hand of Princess Ennigaldi. She had not only collected these items, possibly excavating some herself, but had also taken the trouble to arrange and catalogue them.

Although not all historians are convinced that this is the earliest example of a museum, it is nonetheless a notable early example of descriptive metadata and of metadata designed to support curation. Ennigaldi-Nanna had carefully selected and organized her collection and had created these records to describe its holdings, put them into context and make sense of them.

Ennigaldi-Nanna's cylinder is one of most renowned artefacts of early metadata but it is far from the most ancient. Metadata creation is as old as libraries themselves, although our knowledge of how they catalogued or listed their materials is often sparse as we go further back in time. One claim to be the oldest known library is that of the city-state of Ebla in current-day Syria. Excavations here in the 1970s revealed a collection of several thousand clay tablets from around 2500 BCE, all originally shelved with the first line of each tablet left visible to allow its easy retrieval [2]. Although there is no evidence of a surviving catalogue for this library,

R. Gartner, *Metadata*, DOI 10.1007/978-3-319-40893-4_2

the excavations revealed the existence of a large number of lists, covering everything from gods to birds to professions, the last of which were arranged by rank. These are often taken as an early example of a classification, an attempt to use metadata to arrange the world by constructing a hierarchical view of it.

A more recognizable approach to cataloguing is apparent by the time of the Hittite empire a thousand years later. A prime example of this is the archives of the palace at Hattusa in modern day Turkey, which were found to contain around 30,000 clay tablets, much larger in scale than the collections at Ebla and so in need of more sophisticated cataloguing. Two important advances are in evidence here, although they may not be the first examples of their kind.

One is the use of a *colophon*, literally a summit or top, a short description of contents. This served a function akin to the title page of a modern book, making it easier to identify a work and keep together all of the tablets on which it was carved; this was necessary as many now extended beyond a single tablet. The other innovation seems to have been recognizable catalogues with descriptive metadata or bibliographic information. Lionel Casson, in his *Libraries of the Ancient World*, notes that these included such elements as the title of a work (often its first line), its author's name, a description of its contents and a count of the number of tablets it took up [3].

A further leap in the size of collections, and of consequently more sophisticated cataloguing, can be seen in perhaps the best known library of the ancient world, the Royal Library of Alexandria. Much undocumented legend surrounds this library, but it is clear that it was one of the largest and most prestigious in the world at the time of its construction in the third century BCE and remained so for over 200 years. In the history of library metadata, it has an almost iconic status as the home of one of the most notable early attempts at scholarly bibliography, the *Pinakes* of Kallimachos of Cyrene.

The *Pinakes* (literally 'tablets') were lists of preeminent authors in all branches of literature and details of their works. They were compiled by Kallimachos, a noted poet of the time, while he enjoyed the patronage of the Ptolemy family under whose aegis the library operated (although whether he was in fact the librarian of the institution is generally considered dubious) [4].

A number of fragments of the *Pinakes* survive and show us something of the details that Kallimachos compiled. They include a brief biography of an author's life, their family background, place of birth and death, a list of their works (including their *incipits*, the opening lines by which they were usually identified), and a discussion of any controversies surrounding the attribution of their authorship to any given work (these were very common at the time). This descriptive metadata was also supplemented by its administrative counterpart in the form of a note about the number of lines of text each work took up, information that was designed to make life easier for the librarian [5].

Although the *Pinakes* have sometimes been called the first library catalogue [6], it is probably more accurate to see them as pioneering works in biography and bibliography, rather than as something intended primarily to help users find an item on

the shelf. In addition to the depth they display in putting authors into context, they are also interesting as early attempts to arrange the genres of works in a hierarchy.

Lionel Casson shows how Kallimachos divided writers initially by whether they wrote poetry or prose. He then subdivided poetry into the dramatic and epic poets, the former subdivided again into comedy and tragedy [3]; prose was carved up in a similar way. This was one of the earliest examples of an extensive hierarchical classification scheme; as we shall see later, hierarchy is a overriding principle of much metadata, even in the digital age. One problem with it was how to handle authors who fitted into more than one category (for instance, those who wrote both tragedy and comedy): we don't know whether Kallimachos dealt with this by repeating entries for an author under each category, provided cross-references or whether he merely ignored them. This is a problem that preoccupied many a cataloguer in the centuries to come.

One further innovation from the library of Alexandria in the history of metadata is the first recorded use of an alphabetical ordering of authors' names. This method appears to have been adopted by its first librarian, Zenodotus, although he used the first letter of a name only, filing randomly thereafter [3]. Kallimachos adopted the same approach, stopping after the initial letter: another 400 years would pass before other letters would be used to refine the ordering of names more precisely.

The ancient world clearly had some great minds addressing the problems of organizing knowledge and some of the underlying principles of metadata established then, such as alphabetical ordering and the use of hierarchies for classification, still dominate its practices today. Moving forward a thousand years, we find little had changed in the ways libraries catalogued their holdings. There was nothing on the scale of the library of Alexandria in European libraries until the fifteenth century: most collections in the Middle Ages tended to be associated with abbeys and monasteries whose approaches to cataloguing were often perfunctory.

One notable surviving catalogue of the time is that of Reading Abbey: this takes the form of a beautiful decorated manuscript, compiled in the twelfth century, which lists around 300 books not only from the Abbey's library but also those scattered around its various buildings, including its infirmary. The catalogue occupies only four pages of a more substantial volume which forms an inventory of all the Abbey's possessions, everything from clothing to sacred relics. It is relatively unsophisticated compared to the work of Kallimachos over a millennium earlier: it is divided into basic categories, starting with the Abbey's bibles, then glosses on these, the works of the Church Fathers (such as St Augustine), a limited number of classical works and finally liturgical books such as breviaries [7].

This is very much an inventory rather than an aid to finding the books. No indication of their position on the shelves appears: presumably with such a small collection this was not needed to locate a work. For the purpose of documenting the Abbey's possessions, its primary function, this limited information worked well enough. It would take the arrival of a new invention in the fifteenth century, one that magnified the scale of collections and so made new demands on cataloguing, to give metadata its next push.

The Arrival of Printing

Many changes in the forms and uses of metadata so far were precipitated by changes in the size of libraries; from the small collections at Ebla to the vast library of Alexandria, increased quantities of information required new ways of organizing it. All of these collections shared one constraint on their growth: everything on their shelves was hand-crafted in some way, whether this took the form of carvings on clay tablets or fine calligraphy written on parchment scrolls by meticulous scribes. In the fifteenth century, a technological innovation appeared which increased the amount and diversity of information available enormously, the invention of printing using movable type by the German Johannes Gutenberg.

Gutenberg's invention is often seen as the beginning of mass communication and its effects on diverse areas of society and culture have been well documented by historians, most notably by Elisabeth Eisenstein in her two-volume epic *The Printing Press as an Agent of Change* [8]. Printing changed many facets of information and knowledge and the ways in which they spread: everything was produced more quickly, more cheaply, and with much greater diversity than had been possible before. Such an explosion in information undoubtedly caused many a problem for the librarians who had to cope with its aftermath. One of these was to work out how cataloguing should change to handle the deluge.

One of the first large libraries in the post-Gutenberg era to circulate a printed catalogue of its holdings was the Bodleian Library in Oxford. Dating from 1605, when its collection had already grown to become one of the largest in Europe, it reveals that cataloguing practices had not changed significantly from those of the ancient world. Its main section, numbering 264 pages, is a listing of the library's holdings as they were ordered on the shelves, one page of the volume corresponding to one shelf in the library. These listings are themselves grouped into the four main schools of the University of Oxford at the time, Theology, Jurisprudence, Theology and Arts [9].

The information given here is similar to what Kallimachos compiled a millennium and a half earlier, although without the copious biographical details he provided; this, after all, was designed as a means of making it easier to access the library's collections, not as a work of scholarship. The catalogue lists the author's name, where they come from if they are known by their geographical provenance, the title of the work, its physical size, and its place and date of publication. The publisher is not listed at all, a practice that continued in the Bodleian for a further three centuries, nor are the editions of a work noted. The information is enough to allow a visitor to the library to browse the shelves and identify a work housed there, but little else.

To compensate for these shortcomings, the catalogue includes as a supplement an author index, which points to the page numbers of the main section that record where the works of each writer are shelved. It was not until a later catalogue, published in 1620, that the Bodleian ordered its listings by author for the first time; this reflected a rising awareness of the notion of personal authorship and a recognition that it is through the author's name that most users were coming to identify a work [9].

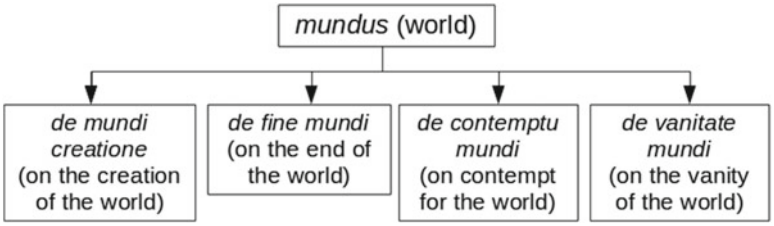


Fig. 2.1 Hierarchical arrangement of entries for *mundus* (world) in James' 1620 catalogue

The Bodleian's Librarian who was responsible for these historic catalogues, Thomas James, recognized their limitations when it came to giving his users more than the most basic access to his collections: they either had to know a work and its author in advance or else were willing to browse the collections by leafing through the thousands of works under each category in the shelf listings. He was, fortunately, more ambitious than this; he particularly wanted, as he stated in the preface to his 1620 catalogue, to 'dig up' the treasures buried in the books in the University's faculties [10]. To do this, he undertook an ambitious set of subject indexes to the collections which occupied some 18 years of his life.

These four indexes, one for each of the faculties, were by far the most comprehensive attempt at the time to categorize knowledge in fine detail. The scale of this undertaking was huge: the theology index alone ran to over 10,000 entries. In its scope and scholarly ambition, James' work was a worthy successor to Kallimachos' *Pinakes*.

As his renowned forebear had done, James used hierarchies to divide up his subjects. The classification for mundus (world) in his theology catalogue takes the form shown in Fig. 2.1.

Unsurprisingly, the categories chosen reflect the pre-occupations of theological thinking of the time which were evidently much concerned with the beginning and end of the world and how contemptible a place it was. The world only merits a small number of subdivisions but other categories are much more finely detailed: *fides* (faith), for instance, has 57 subheadings and *Christus* (Christ) 92 [10].

One problem that James inevitably encountered when trying to divide up knowledge into these neat categories was how to deal with works that did not fit cleanly into a single one. His solution was to duplicate entries extensively. The concept of Holy Communion is covered by the headings *Coena Domini, Communio*, *Eucharistia* and *Missa*, all of which repeat the full list of holdings under each entry [10]. This obviously increased the size of the catalogue substantially: later cataloguing practice would adopt the idea of a main entry for a heading and cross-references ("See…") to point to it.

Great though James' achievement was in these catalogues, they have some shortcomings arising from the fact that they were the work of one man who invented the conventions he applied as he went along. They are full of inconsistencies which became more apparent as they grew in size: anonymous works are sometimes listed under the first word of the title, sometimes not, titles are inconsistently transcribed and so on [9].

The next catalogue, from 1674, made some tentative attempts to address these problems by codifying the rules used within it. Here, for instance, it was decreed that authors known by a pseudonym should be listed under their real name and cross-references provided from the pseudonym to this main entry [11]: this rule persisted in the Bodleian until the 1980s, leading to such quirks as entries for the novelist George Eliot being listed under her real name Mary Anne Evans. Despite the peculiarities they produced, the adoption of cataloguing rules in any form was a great step forward.

The Nineteenth Century: Information Industrialized

The nineteenth century witnessed upheavals in all aspects of life and society, and information was no exception. Mirroring the expansion in material production that was sparked by the Industrial Revolution was a comparable increase in the rate at which information was created and propagated. Unsurprisingly, the ways in which metadata was created and handled had to change to accommodate this new pace and volume.

One significant change that was slowly recognized throughout the library profession was the need for a more solid codification of cataloguing rules than the ad-hoc practices each library had previously adopted. The most influential move in this direction came from the British Museum, which since 1753 had operated as the national library of the United Kingdom. In 1841, the Museum published its *Rules for the Compilation of the Catalogue* devised by its Italian-born librarian, Antonio Panizzi [12]: these guidelines, usually known as the *Ninety-One Rules*, formed the basis of most cataloguing practices for over 150 years.

Panizzi's rules codified guidance on many of the thorny issues which had been treated in idiosyncratic ways before he threw the force of his logical mind at them. Some of these are still in force in contemporary cataloguing practices. His ruling, for instance, that the first named author of a work should be taken as its principal creator unless otherwise indicated (Rule III) is still ingrained in the Anglo-American Cataloging Rules [13] to which most libraries adhere today [14]. Some other quandaries with which Panizzi grappled included the names of rabbis (Rule V), friars and saints (Rule VI), and how to deal with people known only by their first name (Rule VIII).

The mid-nineteenth century also saw the eruption of a schism between the worlds of libraries and archives, one that persists to this day in their very different approaches to metadata. Much of this rift owes its origin to the French archivist Natalis de Wailly, the Head of Administration at the Archives Nationales, who in 1841 proposed that archives should abandon previous ways of arranging their collections, derived from the practices of libraries, and

> assemble the different documents by *fonds*, that is to say, to form a collection of all the documents which originate from a body, an organization, a family, or an individual, and to arrange the different *fonds* according to a certain order. [15]

This may not look like the stuff that deep fissures are made of but it represented a fundamental break from the way librarians treated their collections and metadata. They tended (and still do) to concentrate on the individual item in their holdings: their catalogues describe it in more or less detail but they are not particularly interested in where it fits into the collection as a whole. Archivists, by contrast, have since de Wailly's time followed his principle, known universally by the French term *respect des fonds*, to concentrate on describing the collection and its internal structure: they are less interested in describing each file or piece of paper in an archive than in putting it into context, showing how each part of a collection relates to the whole.

In the archival world, the prime type of metadata record is the *finding aid*, essentially a description of a collection as a single entity. The idea of the finding aid is to help the scholar using the archive locate roughly where they might find material of interest but not to go down to the level of describing each item. Traditionally it includes information on the provenance of the collection, biographical notes on its key players (the collector, perhaps, or the person whose papers it holds), and a description of its arrangement. As so often in the world of metadata, this last component is usually represented as a hierarchy of layers, from '*fonds*' at the top to 'item' at the bottom.

If the nineteenth century gave us one of the great schisms in metadata, it also bequeathed us one of its most ubiquitous ways of dividing up knowledge by subject. Despite the brave attempts of such figures as Thomas James to impose order on the slippery world of the subject catalogue, most approaches to this until the mid-nineteenth century were idiosyncratic and inconsistent. This period saw a rise in science's interest in taxonomy, particularly in chemistry (in the form of the periodic table) and biology (in the systematic application of a classification scheme for all organisms designed by the Swedish zoologist Carl Linnaeus). Unsurprisingly, librarians recognized the need to treat subject taxonomies more systematically: the best-known, and longest lasting, of their attempts to rise to these challenges was the work of the American Melvil Dewey.

Dewey, the head librarian at Columbia University and later the New York State Library during the last two decades of the nineteenth century, published his famous *Dewey Decimal Classification (DDC)* in 1876. Many who have used public libraries will recognize the spine labels on books that carry the DDC numbers for their subjects. Now in its 23rd edition, the DDC divides up the world of knowledge into ten broad categories (Table 2.1).

These then subdivide into great detail within a strict all-encompassing hierarchy. Each subject finds its allotted slot on one of the branches of this tree of topics.

This approach offers one great advantage over the subject classifications encountered earlier. It allows the arrangement of books on the shelves to match their subjects in as much detail as the cataloguer decides to record in their DDC numbers. Previously books tended to be placed in the order in which they were acquired on shelves which reflected very broad categories only (as in the collections documented in the first Bodleian catalogue). Now users could locate a book that precisely reflected what they were looking for by going to its appropriately-numbered place

Table 2.1 Top level categories of the Dewey Decimal Classification (DDC)

000	General works, computer science and information
100	Philosophy and psychology
200	Religion
300	Social sciences
400	Language
500	Pure science
600	Technology
700	Arts & recreation
800	Literature
900	History & geography

on the shelves, and by browsing through its neighbours, find others on the same or closely-related topics. Dewey's system made one of the great pleasures of visiting a library, the serendipitous discovery of new works on a subject by walking around the shelves, much easier than before.

No classification scheme is without its problems, and the DDC is no exception. The rigid hierarchy Dewey uses to divide up knowledge causes problems if a work covers more than one topic: the scheme only allows one number to determine its position on the shelves so some, often arbitrary, decision is needed on its primary subject. A library organized by DDC also requires its users to go through the additional stage of looking up a Dewey number before they can find the books that interest them. In practice this has rarely been a problem: millions of public library users readily take this in their stride and soon get to know the numbers for their favourite subjects. Not for nothing is Dewey's scheme still the most-widely used in the world.

Paralleling these changes in metadata were developments in the physical media used to store it. Until this period library catalogues usually took the form of bound volumes: these were either printed books, such as the Bodleian's early catalogues, or else *guardbooks*, large scrapbooks in which entries were written or printed onto slips of paper and carefully pasted into their appropriate places within these heavy tomes. Neither of these approaches was entirely practical when collections were not static. Updating a guard book with a new entry was always a cumbersome process, particularly if it required a space on a page that was already full and so necessitated the laborious removal and repasting of previous entries. When it came to printed volumes, a full reprint to accommodate changes to collections rarely appeared more frequently than once a decade.

An answer to the inadequacies of the bound catalogue first manifested itself in the revolutionary France of the 1790s. The government had appropriated all ecclesiastical property in 1789 and with it a number of sizable libraries and their collections. Naturally, they wanted to know what they had acquired and so the compilation an inventory was ordered which grew into the idea of a national bibliography. To accomplish this mighty task, a simple set of cataloguing rules was put together, in which it was mandated:-

> to have playing cards on which can be written the name of the work, that of the author when he is known, the place of printing, and the date… All the cards will be brought to a single depot where they will be sorted, placed in bibliographic divisions and sub divisions, and finally divided among different copyists whose united work will form the catalog (quoted in [16]).

The revolutionaries actually borrowed a technique devised two decades earlier by Abbé François Rozier to produce an index to 90 years of the publications of the *Académie des Sciences* of Paris. Their scale of operation was vastly larger, however, encompassing several million volumes which were listed in the space of 3 years. This was not exactly a card catalogue as we know it: it was not intended for library users to find a volume on the shelf but rather as a way of gathering information on libraries which were spread out geographically.

The great advantage of using cards over guardbooks was the ease with which they made multiple entries possible for the same work. One problem had dogged cataloguers for millenia: how to deal with works of more than one subject or which did not fit neatly into a single category? With cards it was simple to create more than one card for each item and provide it with more than one point of access. Another advantage was the speed with which catalogues could be updated: no more pasting and unpasting of slips in guardbooks, now a catalogue could be amended merely by slotting a new card into a drawer.

What is perhaps surprising is that it took so long for these advantages to be recognized and public card catalogues to become widely available. It was not really until the 1870s that they became common in libraries in the UK and USA, partly as a result of it being championed by the seemingly ubiquitous Melvil Dewey [17]. One of the first acts of the newly-formed American Library Association was to agree at its conference in 1877 on a standard size for catalogue cards [18]. This was more significant than it may at first appear: with this standardization came a new uniformity for the cabinets that could hold these cards and new possibilities for sharing metadata on a much grander scale by distributing them.

The Twentieth Century

Much of the history of metadata in the early part of the twentieth century is one in which the great advances of the preceding 100 years were repackaged to allow its creation and dissemination on an ever more industrial scale. But this period was also responsible for the emergence of a more profound change, a new way of looking at metadata itself at a more abstract level: this moved away from the hierarchies that had dominated it for centuries to new models which have come into their own in the digital landscapes of the twenty-first century.

With so much information streaming into the world, and particularly into libraries, it made more sense to centralize the creation of the metadata record to cope with this influx. In the United States, the Library of Congress began in 1901 to

sell catalogue cards, a service that was rapidly taken up by hundreds of libraries because of the time and cost savings it allowed. Now these libraries could simply add their own local finding information, such as their shelfmark, to these cards and file them directly into their standard-sized cabinets, instead of expending their resources on recreating the metadata already produced by the experts in the nation's largest library. This set a trend which has continued for over a century into the electronic era.

The Library of Congress was also responsible for pioneering what would become a cornerstone of librarianship in the twentieth century, the creation of a *union catalogue*. This term describes a catalogue that documents the collections of more than a single library. The *National Union Catalog*, begun in 1901, attempted to catalogue every book of any note in libraries in the USA, a huge undertaking which grew to over 11 million records. Such an enterprise was only possible because of the advances in standardization made in the preceding century. The notion of bringing together metadata on resources that are geographically spread out still informs the electronic gateways to information available today, from contemporary union catalogues to services such as Google which apply the same principles to data itself.

The most significant change to the way we look at the fundamentals of metadata arose once more in the world of libraries, this time in India. The renowned librarian S.R. Ranganathan, one of the most revered practitioners of the profession to this day, came up with a new way of classifying materials which moved away from the strict hierarchies that had dominated it until this time. His *Colon Classification*, first published in 1933, attempted to get round the problem of systems such as the Dewey Decimal Classification which required that a book was slotted into a single place in its hierarchy despite the likelihood, in almost every case, that several others could summarize its subjects with equal validity.

He did this by defining small units of topics which could be combined flexibly to form more complex compound subjects. These units he termed *facets* and the schemes in which they are used, of which Ranganathan's was the first of any significance, are called facetted classifications. This approach was a great move forward in terms of meeting the challenges of information that is constantly in flux as human knowledge advances. Its drawback when used in libraries is that it produces complex shelfmarks which are much harder for their staff and readers to comprehend than the simple digits of a Dewey number. But in the digital world, where these constraints no longer apply, faceted classification has come into its own: here the capacity to combine facets at will can easily be realized. As we shall see later in this book, faceted approaches to finding data are now everywhere. Ranganathan's innovation is perhaps one of the most forward-looking legacies of the pre-digital metadata world to its electronic descendants.

References

1. Woolley, L., & Moorey, P. R. S. (1982). *Ur "of the Chaldees": A revised and updated edition of Sir Leonard Woolley's Excavations at Ur*. Ithaca: Cornell University Press.

2. Wellisch, H. H. (1981). Ebla: The world's oldest library. *The Journal of Library History, 16*, 488–500.
3. Casson, L. (2001). *Libraries in the ancient world*. New Haven: Yale University Press.
4. Fraser, P. M. (1972). *Ptolemaic Alexandria: Notes*. Oxford: Clarendon.
5. Witty, F. J. (1958). The Pinakes of Callimachus. *The Library Quarterly, 28*, 132–136.
6. Blum, R. (1991). *Kallimachos: The Alexandrian Library and the origins of bibliography*. Madison: University of Wisconsin Press.
7. Westwell, C. (2012). Found in a Bricked-Up Chamber: The Reading Abbey Library Catalogue – Medieval manuscripts blog. http://britishlibrary.typepad.co.uk/digitisedmanuscripts/2012/11/found-in-a-bricked-up-chamber-the-reading-abbey-library-catalogue.html. Accessed 31 July 2015.
8. Eisenstein, E. L. (1979). *The printing press as an agent of change: Communications and cultural transformations in early modern Europe*. Cambridge: Cambridge University Press.
9. James, T. (1986). *The first printed catalogue of the Bodleian Library, 1605: A facsimile*. Oxford: Clarendon.
10. Wheeler, G. W. (1928). *The earliest catalogues of the Bodleian library*. Oxford: Oxford University Press.
11. Bakewell, K. G. B. (1972). *A manual of cataloguing practice*. Oxford: Pergamon Press.
12. Panizzi, A. (1841). *Rules for the compilation of the catalogue*. London: British Museum. Department of Printed Books.
13. American Library Association. (2012). Anglo-American Cataloging Rules homepage. http://www.aacr2.org/. Accessed 28 June 2013.
14. Lehnus, D. J. (1972). *A Comparison of Panizzi's 91 Rules and the AACR of 1967*. Urbana Champaign: University of Illinois Graduate School of Library Science.
15. Bartlett, N. (1992). Respect des fonds: The origins of the modern archival principle of provenance. *Primary Sources and Original Works, 1*, 107–115.
16. Hopkins, J. (1992). The 1791 French cataloging code and the origins of the card catalog. *Libraries and Culture, 27*, 378–404.
17. Sharp, H. A. (1937). *Cataloguing: A textbook for use in libraries*. London: Grafton.
18. Krajewski, M. (2011). *Paper machines: About cards & catalogs, 1548–1929*. Cambridge, MA: MIT Press.

Chapter 3
Metadata Becomes Digital

Well before anything that we would now recognize as a computer appeared, some metadata had moved away from its written and printed forms into something resembling the digital format in which most of it is stored today. To see an early example of this, we could go back to eighteenth century France and its flourishing textile industry. Here we find weavers generating complex patterns in the cloth they manufactured by selectively ignoring some of the vertical threads they raised up when passing across the shuttle containing the horizontal yarn. This was a complicated business and easy to mess up with results that could be financially ruinous.

Fortunately, a clever invention eliminated the possibility of human error ruining the woven fabrics that were being produced so laboriously. Some looms used perforated paper rolls to control which threads were lifted up and which left in place to create these patterns. By the 1720s these delicate rolls had been replaced by more robust punched cards, which reached great levels of sophistication by the early nineteenth century. Figure 3.1 shows a *Jacquard loom,* the invention of the French weaver Joseph Marie Jacquard and an impressive example of the technology from that time.

We are looking here at an early, but by no means primitive, way of automating metadata. The data in this case is the woven fabric and its elaborate patterning. The metadata is an abstraction of the intended patterns, a template from which the data can be constructed when processed by mechanisms that can decipher the holes in the cards. Metadata has moved away from a human-readable, analogue encrustation of ink on paper or parchment. It is now reduced to on-and-off patterns, akin to a binary code, and processed by machines.

In the twentieth century, electronics eventually replaced the mechanics of the punch card and metadata found its home in the magnetic storage media on which most of it resides today. With the new media came the need for new formats, often to do the same job as before but in a way that new methods for information processing could use more effectively.

R. Gartner, *Metadata*, DOI 10.1007/978-3-319-40893-4_3

Fig. 3.1 Jacquard loom (1801) (Photograph by George H. Williams)

Enter MARC

As so often in the history of metadata, librarians were once again to prove themselves pioneers, here to take up the challenges brought about by the move to the digital. By the 1960s, the Library of Congress, long a supplier of printed catalogue cards to thousands of libraries and the custodian of the first union catalogue, had come to recognize the necessity of a new format for sharing cataloguing records electronically. The outcome of this forward thinking was the introduction of one of the most pervasive and long-lasting metadata standards, known to this day as the *MARC* format.

MARC stands for *MAchine Readable Cataloging*, a prosaic but wholly accurate summation of what it is designed to do. It was intended to produce catalogue records capable of processing by computer, not only for searching (its primary function) but also for editing, updating, transferring between systems and sharing. In a word, the

```
001       000022605
003       UkLU-K
008       881221r19861970en-a----r-----001-0-eng-d
015       $a b8709014
020       $a 0140552049
041 1     $a engfre
100 2     $a Lévi-Strauss, Claude, $d 1908-
240 13    $a Le Cru et le Cruit. $l English
245 14    The raw and the cooked / $c Claude Levi-
        Strauss / $c translated from the French by John
        and Doreen Weightman
260       $a Harmondsworth : $b Penguin , $c 1986,
        c1969
300       $a 369p : $b ill (pbk) ; $c 20cm
490 0     $a Peregrine books
490 1     $a Introduction to a science of mythology ;
        $v 1
534       $p Originally published: New York: Harper &
        Row, 1969 ; London : Cape, 1970. - Translation
        of: Le Cru et le Cuit
504       $a Bibliography: p343-352. - Includes index
650 00    $a Indians $x Religion and mythology
650 00    $a Mythology $x Case studies
800 2     $a Lévi-Strauss $h Claude $t Introduction to
        a science of mythology $v 1
```

Fig. 3.2 MARC record (truncated) (Courtesy of King's College London Library)

format enabled metadata to become *interoperable*, a malleable, flexible object which would allow disparate catalogues to communicate and work together automatically. It was a bold aim but the format, still in use half a century later, has proved more than capable of realizing it.

The MARC format was the result of a collaboration between librarianship and computer science. Its 'author' was one of the first computer programmers, Henriette Avram, who started her career with the same National Security Agency who have more recently taken such an interest in all our metadata. Engaged by the Library of Congress to work out how catalogue records could be rendered machine-processable, she had a very large reproduction of a catalogue card mounted on her wall and, with two colleagues, analyzed every part of the record in minute detail [1]. From this analysis, she derived the fields which forms the MARC record as we know it today.

What Avram did was to dismember a bibliographic record and reduce each field to three components, its name, instructions on how to handle it and any logical subdivisions of its contents. The result was a record which looks like the example in Fig. 3.2.

Instead of textual labels for each field, Avram decided to use three-digit numbers which are usually called MARC tags (displayed at the beginning of each line in this example). This made sense for many reasons: they are not specific to a given language, they are easier for computers to process and they are less prone to error (there is no possibility of getting upper- and lower-case letters mixed up, for instance, something that could easily throw machine processing out of kilter).

Trained cataloguers know these numbers off by heart – 100 to them will always be the *Main Entry – Personal Name*, 245 the *Title Statement* and 300 the *Physical Description*.

The digits that appear after some of these three-digit numbers are instructions on how a field's content should be processed when the record is displayed. To take one example, the digits *14* after the *245* for the *Title Statement* provide two directions: the first indicates that this title should be used as an 'added entry', a secondary way to find this item in the catalogue in addition to the author's name; the second indicates that when compiling an alphabetical index of titles, the first four letters or spaces of the title ("The") should be ignored to avoid listing together all entries that begin with the English definite article.

The final feature of the MARC record that Avram introduced is the use of subsections (known as subfields) within most of its principal fields. The 260 field in this example, for instance, which provides information on the *Imprint* of the book (its publication details), has separate subfields for the place it was published (marked by the tag **$a**), the publisher's name (**$b**) and the date of publication (**$c**). Using these means that these smaller components can be used in their own right when necessary (for instance, to compile a list of books by the same publisher), but that the record as a whole uses only a relatively small set of fields and is easier to handle as a result.

One seemingly incongruous feature that the eagle-eyed computer scientist will spot is how much redundant punctuation there is in this record. The same 260 (Imprint) field contains a colon before the **$b** for the publisher's name and a comma before the **$c** for the date of publication. These are not needed at all: you can easily get the system displaying the record to put them in by using the subfield tags for guidance as to what should go where. Why then are they included?

The answer is that the MARC record owes its origins to the card-based catalogues that it superseded and emulates many of their practices. A catalogue card for this same book would take something of the form in Fig. 3.3.

The card follows strict formatting conventions which had long become established as standard practice in most libraries: each component is defined not by tags as in the MARC format but by its placement in the record and the punctuation symbols that precede or follow it. The publisher's name here is preceded by a colon and a space and followed by a comma and a space. This might seem unduly pedantic, but it does mean that any record in any language in the world that followed these conventions would be readily understandable: the publication information between a colon and a comma would always show the publisher's name and nothing else. So set in stone are these conventions that they are usually included in a MARC record to this day.

Another quirky feature of MARC that betrays its origins in earlier catalogues is the way in which it has separate fields for *main* and *added* entries. In a traditional catalogue, one record would usually act as the primary entry point for a book: this would be filed under the field it was assumed most users would tend to look under, usually the author's name. This would often be the only full record for the book. Other shorter ones might point to it, perhaps to record alternative forms of the

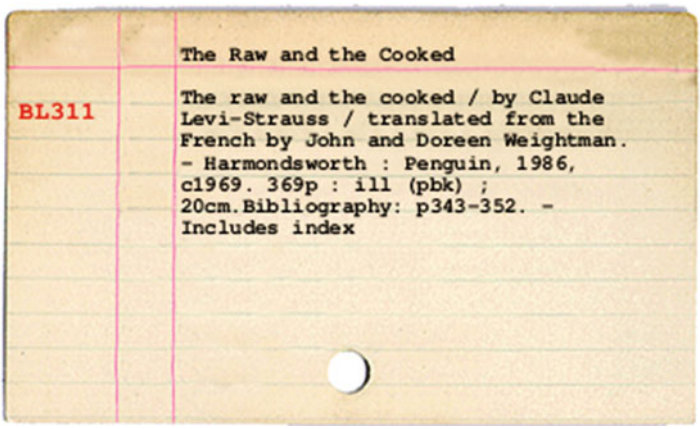


Fig. 3.3 Sample catalogue card (Generated by http://www.blyberg.net/card-generator/)

author's name, but these would most definitely be secondary entries, often with much sparser information than their primary counterpart.

These practices hearken back to the earlier days of catalogues housed in large bound volumes, where inserting records was often a cumbersome and time-consuming process and so denoting one as the primary entry point minimized the laborious work involved. It was less troublesome to duplicate full records in card catalogues but the practice continued nevertheless. In a computerized catalogue, any part of a record can be used as an entry point and so it makes no sense to promote one field to the august status of the main entry. Nonetheless, convention persists and cataloguers still follow complicated rules to work out which part of a record should be honoured in this way.

These quirks should not be surprising as metadata inevitably reflects the community it comes from and so tends to follow its conventions. Nor should they distract from the immense step forward that the MARC record represents. It was one of the earliest attempts to make metadata truly interoperable and in that it has been an enormous and enduring success. The world of libraries has certainly been transformed out of all recognition by its adoption. One of most groundbreaking advances it made possible was the creation of automated union catalogues, successors to the Library of Congress' National Union Catalog but on much grander scales.

A service known as *WorldCat* [2] is a fine example of this. This brings together the catalogues of 72,000 libraries and allows them be searched in one go, more than 330 million records. All of this is only possible because these libraries have adopted and carefully applied the MARC standard. What Kallimachos of Cyrene would have made of this we can scarcely imagine, although in many ways it represents the summation of the work initiated by him and his fellow ancient cataloguers. Certainly the achievements of Henriette Avram, one of the unsung heroes of the information world, should be considered on a par with these now legendary pioneers.

Enter the Internet

While the MARC standard was embedding itself in the library world, profound technological changes were beginning to emerge in the way information was moved around the world. A method known as *packet switching* was developed in which data is collected into small blocks which can then be transmitted efficiently over digital communication networks. This new way of transferring information naturally interested the United States Department of Defense who in the 1960s funded the creation of a network based on it. From their perspective the great strength of this network, which became known as ARPANET, was that it had no centre, no single machine on which it ran: there could be no way in which it could be knocked out in its entirety by a single blow from the enemy.

ARPANET was the foundation of what we now know as the Internet. This network-of-networks expanded greatly in the 1980s when universities and later commercial providers gained access to it. They could do this because they all adopted the same *protocol,* the same set of rules for packaging, sending and receiving data, which had first been developed by ARPANET and was known, unsurprisingly, as the *Internet Protocol Suite*. The Internet was a flourishing and rapidly expanding medium well before 1989 when Tim Berners-Lee came up with his simple idea for sharing documents over the Internet which we now know as the *World Wide Web (WWW)* or often simply as *the Web*.

The rest, as we know, is history and well-documented history at that. For metadata, the Web presents a number of profound and often contradictory challenges. So much data is out there, and so much is added to it every second, that it might seem more important than ever to have good quality metadata to find it. Managing all of this data might also require huge quantities of administrative metadata, technical and otherwise, to curate and look after it. How can we generate and maintain metadata in such huge quantities? Perhaps we don't need it at all. Google and other search engines, which allow us to search the content of almost anything on the Internet directly, offer a challenge to the traditional notion that we need metadata to make sense of data. If we can search the contents of a library's collections directly, why do we need its catalogue? If we can find what we want on the Internet without the middleman of metadata, why bother with the time and expense of creating it?

The challenge of Google and its content-based retrieval to metadata will be examined in a later chapter. In reality, the Internet and the Web have not killed off metadata or the need for it: instead, it has made necessary new types and new standards. When we look at how it has developed in the two decades that have elapsed since Tim Berners-Lee wrote his first Web document, we can see that not only is metadata as important as ever but that it is still recognizably attempting to do the same as it has always done. The same fundamental principles it has followed throughout history still apply to it today.

A MARC Standard for the Internet: Dublin Core

The early days of the Internet, particularly before the advent of Google and other search engines in the latter part of the 1990s, saw it being thought of as something of gigantic library which needed cataloguing in much the same way as any other. Some of the early attempts to do this now seem quaint to say the least. A brave one in 1994 tried to catalogue the Internet in book form, akin to the bound volumes of the seventeenth-century Bodleian Library [3] Others more realistically embraced the models of electronic catalogues. The Bodleian Library once again pioneered developments in metadata when it launched an early portal to the Internet in 1992 [4], using an ad-hoc scheme to describe the small number of resources to which it provided access. Later, much larger catalogues such as Yahoo (launched in 1995) also tended to make up their own conventions.

A consensus was reached that a more rational and standardized way of describing resources on the Internet, particularly those on the Web, was necessary, something akin to the MARC record for books on a library shelf. Like the MARC record, this would allow online resources to be discovered, records to be shared and large portals to these resources to be built up which would be analogous to the union catalogues of the library world. The problem with trying to emulate MARC was that objects on the Internet are much more diverse than those found in a library: this makes it much harder to decide on a set of elements that can handle everything one might wish to say about this jumble of data. Henriette Avram had long-established and highly-focussed practices to work on when putting together the MARC standard: where to start in a medium only a few years old and which was changing by the day?

Taking on this challenge was a brave set of experts from the worlds of libraries, computer science, museums, online information services and archives who met in Dublin, Ohio in early 1995. Instead of attempting to emulate the complex MARC record, and potentially create a huge set of elements to cope with anything that might be found on the Web, they chose instead to aim for simplicity and pare things down to a minimum. Their intention was to identify a core set of elements that could be applied to any object, physical or digital: the standard they came up with acquired the name *Dublin Core*.

Dublin Core in its simplest form consists of a mere 15 elements, each of which is defined broadly and in a way that should be comprehensible to anyone, not just an expert in metadata. These elements, which make up what is usually called *Simple Dublin Core*, are shown in Fig. 3.4.

The majority of these do not need much in the way of explanation. Most people readily understand the concept of a title, although working out what exactly is the title of something like a webpage can be rather more difficult than grasping the concept in itself. A few do usually need some elaboration. The difference between

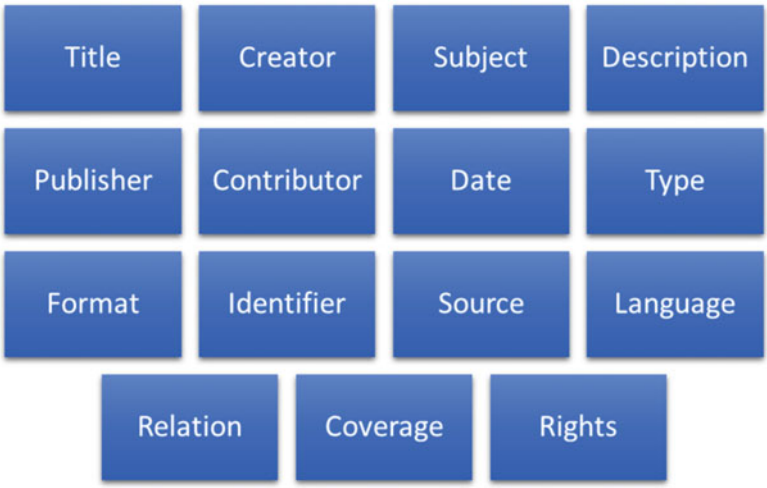


Fig. 3.4 The 15 elements of Simple Dublin Core

Creator and *Contributor* often causes confusion: the creator element is intended to record the person or body principally responsible for the intellectual content of an item while the contributor notes the people or bodies who have supplemented the work of the creator in bringing it into being. The difference between the two can be a little obscure and their boundaries can be fuzzy: is, for instance, the director of a movie the only person who should be in the creator field (auteur theorists might thing so) or should other creatives such as the screenwriter or producer go here as well?

The simplicity of this set of elements is a great strength of Dublin Core. None is mandatory and as many or as few as are needed can be used to describe something. This 'something' can be anything at all: Dublin Core records can cover the physical (a book, a dress, a lampshade) and the digital (digital versions of these or something that is 'born-digital', not a copy of anything in the physical world). Simple Dublin Core has more than proved its worth for this purpose and can be found in all corners of the Web.

Despite this great success, Simple Dublin Core presents a potential problem for anything that requires more than the most basic metadata. Its elements are so broadly drawn that if we find two records using the same element we cannot be sure they mean the same thing. Does one website's notion of a *Creator* mean the same as another's? And won't those trying to find something on the Web want something more specific than an element as wide as this to allow them to differentiate the creative roles associated with it? If we find both Alfred Hitchcock (the director) and David O. Selznick (the producer) listed under *Creator* for a film such as *Rebecca* (1940), wouldn't we want to know in more detail who did what?

Various remedies have been devised to alleviate this problem and make Dublin Core more precise when necessary. One that was formulated shortly after

the standard was first published is known as *Qualified Dublin Core*: as its name implies, this is a method for adding qualifiers, or supplementary tags, to Dublin Core elements to render them more specific. The *Creator* element may be qualified to indicate that it refers to an author of a book in this way:-

```
creator.author
```

Anything can be used as a qualifier, although to make things more consistent the body that maintains Dublin Core publishes a list of those that it particularly recommends. The clever part of this way of qualifying an element is that it still allows it to be understood by a system which does not comprehend the qualifier itself: simply removing everything after the dot (.) in the element name leaves us with the unadorned Simple Dublin Core *Creator*. Even if the system reading this record has no idea what an author is, it will still know that this element refers to the person or body primarily responsible for the intellectual content of the object. In this way we can hedge our bets: we can be as specific as we wish on our own system but be sure that our records can still be understood, albeit in a more basic form, on any system that can recognize Dublin Core.

Dublin Core has undoubtedly become the *lingua franca* for descriptive metadata on the Internet. Simple or Qualified Dublin Core form the basis of millions of websites and underlie the architectures of many a digital library or repository. Its elements have also been incorporated into other schemes that supplement this small set with extra elements to create larger, more complex and more specialized standards; these include one, the widely-used DCTerms, that is produced by the Dublin Core team themselves [5]. Despite its many detractors, it has undoubtedly fulfilled its initial remit of providing a generic, easily implemented way of encoding and sharing metadata on resources in the physical and digital worlds.

Metadata Behind the Scenes

The history of metadata covered so far in this book has mainly concentrated on its descriptive form, but the advent of digital information has made its administrative counterpart equally important. Very few systems that serve up digital objects, anything from ebooks to audio to video, could operate without an abundant set of metadata operating behind the scenes. As the technology of the digital media has developed, so have the metadata standards needed to make it work.

One of the first types of object that became popular on the Internet, particularly when such early graphical browsers as *Mosaic* and *Netscape* appeared in the 1990s, was the still image. New compression techniques, such as the still widely-used JPEG (Joint Photographic Experts Group) format, allowed the file sizes of images to be reduced dramatically, so making it feasible for them to travel down the slow lines that connected computers in the early days of the Web. To enable systems to make sense of digital images requires metadata that describes their technical

make-up; sure enough, there is a standard for this put together by experts in the field and known by the unpromising acronym *MIX (Metadata for Images in XML Schema)* [6].

MIX has plenty of siblings covering other types of digital objects. Video has its own technical metadata standards, audio its own, and even text, seemingly one of the simplest types of digital object conceivable, has several as well. Not all of these have achieved the dominance of the MARC standard in their respective domains: technical metadata for video, for instance, can be found in more than one scheme and the choice of which to use often depends on the community employing it. If we were looking for cosmic analogies, we might think of these competing standards as the dwarf planets of the metadata solar system: they have not cleared their orbits of others of their kind and find themselves in a similar predicament to poor, demoted Pluto when compared to their all-powerful counterparts such as MARC. The metadata universe, like the physical, can be a messy place.

Old Divisions Cut Deep

The media for storing metadata may have changed out of all recognition since the pre-computer age, but something that has not is the divisions that have their roots in that now-distant era. We have already seen that the MARC standard bears the imprint of the catalogue card that it replaced. That card, with its emphasis on the single item in the library, itself reflected an approach to metadata from which the archival world had long declared independence since that historical moment when Natalis de Wailly proposed the concept of the *respect des fonds*.

This division lives on in the very different approaches to metadata that persist in the world of archives to this day. Instead of adopting MARC, archivists have created their own metadata standard for describing collections, known as the *Encoded Archival Description (EAD)* [7]. EAD takes the structure of a traditional printed finding aid and translates it into a machine-readable form. As one would expect to find in its paper antecedent, an EAD file usually contains sections on the history of an archive, biographical information on the people it covers, narrative descriptions of its contents, and so on. The structure of the collection itself is usually described in a hierarchy from *fonds* to item, exactly as would be done in its counterpart from the nineteenth century onwards. All of this is recorded in one, often very large and cumbersome, electronic file.

Emulating the finding aid in this way certainly allows archivists and their users to stay within their comfort zone but does not exploit the potential of the electronic medium to create more flexible ways of searching archival collections. EAD files tend to be monolithic affairs, less receptive to forming the large union catalogues common in the MARC world. Some brave attempts have been made to allow some merged access to EAD records: a notable one was conducted by the UK National Archives who created a project named *Access to Archives* [8] which by dint of sheer ingenuity managed to do this for thousands of records. Services such as this are

rather more primitive than their cousins in the library world; they are generally limited to fairly simple searches on a small number of fields because EAD does not contain the type of metadata that computers can readily process in sophisticated ways.

Another division, not quite deep enough to be described as a fissure but certainly a crack, lies between the printed books cataloguer who inhabits the world of MARC and the cataloguer of manuscripts, particularly of the historical variety. Those who document such manuscripts often consider themselves more than mere cataloguers; they are supposedly scholars whose descriptions are works of erudite research in their own right. In the pre-digital age, their catalogue records were often lengthy expositions, recording in considerable detail anything from the history of a manuscript to its physical make-up, details of the scripts in which it is written and extensive descriptions of its intellectual contents.

In the digital age, cataloguers have attempted to create machine-readable standards for these which have tended to follow the approach of EAD and replicate the format and content of the records produced by their forebears. The most notable of these is a metadata add-on to the *Text Encoding Initiative*, a widely used standard for encoding electronic texts [9]. Although it makes some attempts to pull out the more data-like components of a record, most of it consists of discursive prose descriptions of the type found in traditional manuscript catalogues. As a result, there are no usable union manuscript catalogues of the type found in the world of printed books, just a small number of local catalogues for individual collections.

There is a certain irony in the arrival of a technology that allows information to be shared amongst its counterparts with an ease that could only have been dreamt of by earlier generations coinciding with the persistence of long-entrenched schisms. But, as we saw in Chap. 1, metadata is a human construct and so will always reflect its provenance: divisions between communities and within them will find their analogues in the metadata they devise and use.

Standards Everywhere

For those who predicted that the computer age would see the demise of metadata, a salutary reminder of how misplaced this notion is can be seen in the diagram in Fig. 3.5 by the noted digital librarian, Jenn Riley. Entitled *Seeing Standards*, it offers a map of the world of metadata standards at the time of its compilation in 2010. Contained within its ovals are the acronyms for over one hundred of these; each represents something over which luminaries from their respective communities have deliberated in detail to distill what they regard as the essence of their collective knowledge into a finely-honed set of metadata rules and instructions.

Riley's diagram shows, as we would expect, that there are different metadata standards for different types of data (labelled *Domain* in her representation). There are also different standards for the diverse functions and purposes that metadata can fulfil, which again is no surprise. What is interesting is how segregated standards are

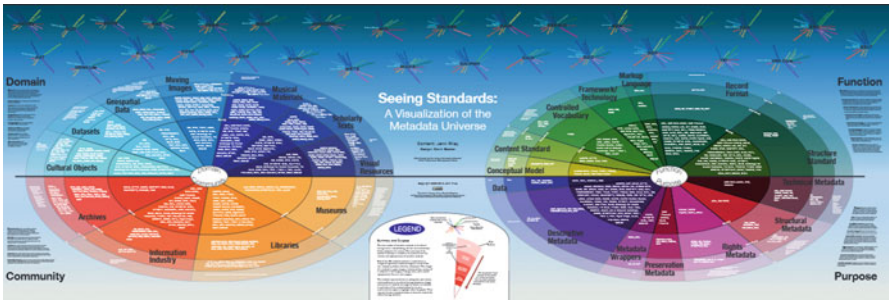


Fig. 3.5 Seeing Standards: a Visualization of the Metadata Universe by Jenn Riley (Reproduced by permission of the author)

by community (shown in the bottom left of the diagram): museums, libraries, the information industry and archives all have their own standards and little overlap is permitted, or even desired, between them. Far from technology bringing together metadata to fulfill visions of a turning the Internet into a single, vast library of information and knowledge, it appears that metadata is fragmenting into factions as resolutely as at any time in its history.

Standards are complex things: the specification for a single one is apt to run into hundreds of pages and to take many months to learn. It is not surprising that once the effort has been made to get to grips with the standard that predominates in a given community few will want to cross the line into the unknown territory of others. It is not peculiar to see why metadata practices have become as acutely entrenched now as they have throughout history.

Not for nothing did the noted computer scientist Andrew Tanenbaum joke in 2003 "The nice thing about standards is that you have so many to choose from" [10]. There is certainly no sense of an end point in this historical survey of metadata, no feeling that we have moved towards a single, universal approach to organizing our data, information, knowledge, understanding and wisdom. We are swimming in a torrent of metadata as resolutely as we have at any point in its history: the media that hold it have changed out of all recognition, human nature has not.

References

1. Pattie, L. W. (1998). Henriette Davidson Avram, the great legacy. *Cataloging and Classification Quarterly, 25*, 67–81.
2. OCLC. (2015). WorldCat.org: The World's Largest Library Catalog. http://www.worldcat.org/. Accessed 5 Apr 2013.
3. Abbott, T. (1994). *Internet World's on internet 94: An international guide to electronic journals, newsletters, texts, discussion lists, and other resources on the internet*. London: Mecklermedia.
4. Gartner, R. (1994). Libraries of the future: Bodleian's BARD online access. *Computers in Libraries, 14*, 53–58.

5. Dublin Core Metadata Initiative. (2016). DCMI Metadata Terms. http://dublincore.org/documents/dcmi-terms/. Accessed 9 Mar 2016.
6. Library of Congress. (2011). Metadata for images in XML standard (MIX). http://www.loc.gov/standards/mix/. Accessed 28 Jan 2010.
7. Library of Congress. (2010). EAD: Encoded Archival Description version 2002 official site. http://www.loc.gov/ead/. Accessed 28 Jan 2010.
8. The National Archives. (2015). Discovery | The National Archives. http://discovery.nationalarchives.gov.uk/. Accessed 28 Aug 2015.
9. TEI Consortium. (2013). TEI element msDesc (manuscript description). http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-msDesc.html. Accessed 10 Dec 2013.
10. Tanenbaum, A. S. (2003). *Computer networks*. Upper Saddle River: Prentice Hall PTR.

Chapter 4
Metadata as Ideology

As has been clear in the last three chapters, metadata is very much a human creation and bears the imprint of its progenitors in the form it takes and the way in which it is deployed. It presents a view of the world which is inevitably subjective and selective in what it chooses to describe and how it attempts to do this. Because metadata is not objective but is an expression of what some philosophers denote by the expressive German term *Weltanschauung* (literally a world outlook), it immediately enters the realm of ideology. It is the relationship of metadata to ideology, consciously and unconsciously, that is the subject of this chapter.

Ideology is a potent term and often used in a derogatory sense to evoke images of authoritarian or totalitarian societies brainwashing their subjects into subservience. But originally it had a decidedly more neutral tone to it. It emerged without any disparaging connotations in early post-Revolutionary France where it was coined by the philosopher Antoine Destutt de Tracy; he used it to describe his "science of ideas" (its literal meaning), a view of human thought as an activity of the nervous system, a combination of sensations without which there would be no such thing as knowledge [1]. The term *ideologue* took on a derogatory tone when Napoleon took umbrage at the influence of the followers of this philosophy whom he blamed (or rather scapegoated) for many of the countries ills [2]. Since then it has struggled to shake off this negative connotation.

Such a charged term has naturally been subject to a variety of definitions, many of which no doubt reflect unconsciously the ideology of their creators. One of the most concise and pertinent for the purpose of this discussion is given by the musicologist Nicholas Cook who, in his *Music: a very short introduction*, defines it as:-

> a system of beliefs which is transparent, which represents itself as 'the way things are' [3].

This seems an accurate summary of what many who are not using the word as a political weapon would consider ideology. It is a set of beliefs that claims to give an unbiased, crystal-clear picture of the world as it is. Cook points out, specifically in reference to the Cold War rhetoric of the 1980s, that it was then "received wisdom that ideology was what the other guy had…Capitalist democracy wasn't an ideology. It was just the way things were" [3].

R. Gartner, *Metadata*, DOI 10.1007/978-3-319-40893-4_4

From this definition it might seem almost a truism that metadata is at least partly ideological. This ideological component stems from the obvious fact, pointed out by philosophers for centuries, that our knowledge of the universe is not and can never be perfect. There are always gaps in it which we inevitably fill in with what we believe to be true. We may be honest about these gaps and acknowledge them when we present our view of the world, as all good academic research should do. But in everyday life we rarely have the time and energy to acknowledge every lacuna in our knowledge and so have to join the dots that we know by making reasonable, if not immediately verifiable, assumptions about what lies in-between. We assume that these are "the way things are" and proceed on that basis as we look at the world.

Metadata is what draws the lines that join the dots of knowledge and so it has to deal with these uncertainties; like its human progenitors, it is unlikely to expose every break in these lines to detailed scrutiny. It generally offers a view of the world which it represents as a transparent reflection of "the way things are", a filter of this presumed reality perhaps (to aid comprehension and to render it usable), but not one that willfully attempts to distort actuality, merely present it in a more digestible form. There is nothing inherently dishonest about this, it is a sincere attempt to express a picture of the world which is genuinely perceived as accurate.

Like all truisms, this is a pretty banal and somewhat pedantic conclusion. It tells us little more than that cataloguers, or others who create metadata in any form, are as human and imperfect as the rest of us. More interesting is to examine how metadata can and has been used to make explicit ideological points or to help the creation of world views within which particular ideologies can thrive. Here metadata is used to do more than cover over the cracks of imperfect knowledge; instead it expresses a view of the universe that is underlain by a belief system which does not acknowledge these gaps.

Mapping the Universe: Cartographic Ideology

In Chap. 1, we looked at an antique globe as an example of metadata and its human origins. The world as we know it does not have such features as lines of longitude and latitude or the names of its oceans imprinted on it but the globe adds these to aid our comprehension of the enormous landmass under our feet. The intention of its makers is laudable and unlikely to be making an ideological point, but those who map the world and the cosmos as a whole may be putting more into their metadata abstractions than a simple reflection of physical reality.

Early views of the universe are far from our current conception of it and their representations now seem quaint at the very least. The picture of the cosmos that held sway in the Western world for many centuries is represented in a beautiful drawing by the seventeenth-century Dutch artist Johannes van Loon (Fig. 4.1).

This shows the universe with the Earth at its centre and all of the heavenly bodies, including the sun, forming a stately procession around it in perfect circles. This

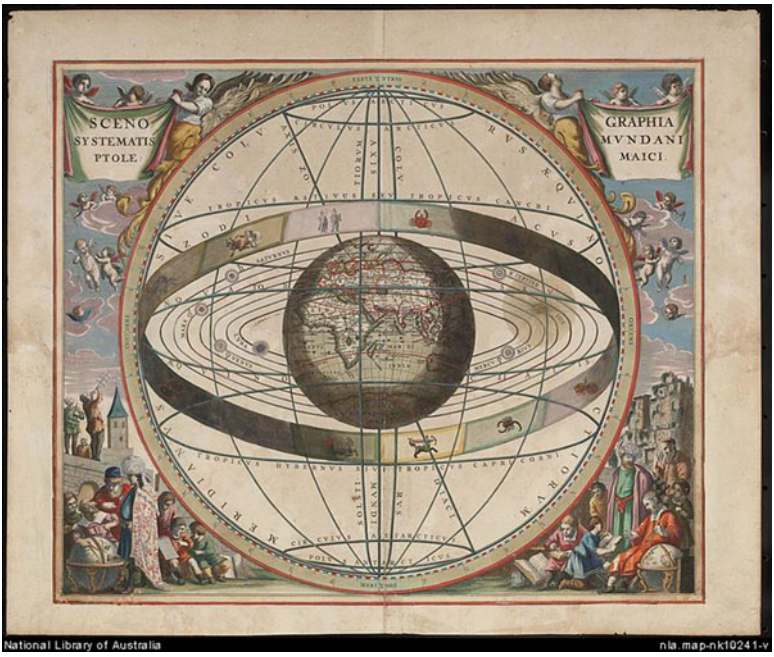


Fig. 4.1 Scenographia systematis mvndani Ptolemaici (Johannes van Loon) (National Library of Australia: nla.map-nk10241)

cosmological model is often referred to as Ptolomaic, after the Alexandrian astronomer Claudius Ptolemy whose *magnum opus*, the cosmological treatise commonly known as the *Almagest,* modelled the universe in this way. At the time Ptolemy drew up his theories he was undoubtedly expressing what then seemed the most logical and consistent way of describing the motions of the heavenly bodies that science allowed.

By the time of this drawing, approximately 1660, this geocentric view of the universe had been challenged over a hundred years earlier by Copernicus in his posthumously-published *De revolutionibus orbium coelestium* (On the Revolutions of the Celestial Spheres). This work relegated the Earth to an ignominious place as the third planet from the Sun (though the Sun itself remained at the centre of the universe rather than in the inconspicuous place we now know it occupies). Less than 30 years before this drawing was made, Galileo had been tried for heresy and threatened with torture and possibly the stake for his heliocentric opinions, a fate which had been already suffered in 1600 by another proponent of Copernicus' model, the Dominican friar and philosopher Giordano Bruno.

For much of history, any map of the universe was metadata that moved from expressing an imperfect science as best it could to one that expressed an ideological belief; a belief, certainly, but one which, as ideologies do, claimed to be presenting

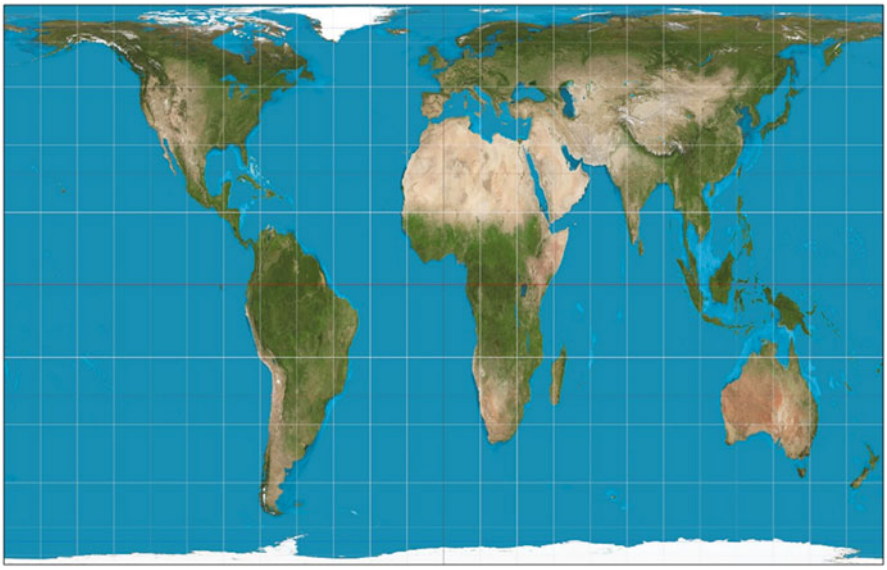


Fig. 4.2 A view of the world (from NASA data) in Gall-Peters projection (Daniel R. Strebe, 2011)

a clear window into reality. For some, the wrong metadata, expressing the wrong world view, could mean extinction.

Although no cartographer today need fear the stake for their work, mapping the world is not entirely free of the influence of ideology. One area where this still rears its head is in the projections needed to render the curved surface of the world onto a flat piece of paper. The venerable Mercator projection, designed in the sixteenth century and still the basis of many online mapping services (including Google Maps) found itself out of favour in last part of the twentieth century partly because it exaggerates regions around the poles and renders those nearer the equator much smaller than their true area. This concern was not only a matter of geographic accuracy but also one of the apparent denigration of the importance of those developing countries that occupy the equatorial and tropical regions.

The answer was a new projection that preserved the relative sizes of the world's landmasses by stretching the equatorial regions north-south and compressing those nearer the poles: this produced the famous Gall-Peters projection (Fig. 4.2) which formed the basis of many wall maps in the late twentieth century.

Metadata here certainly plays an ideological function, albeit one that acts as a corrective to the ideological content (conscious or otherwise) of previous ways of mapping the world. Certainly, the idea of the projection as a way of reasserting the dignity and value of the developing world was one of the major selling points of the Gall-Peters map when it first appeared.

Describing the World: Terminology as Ideology

The terminology used in metadata is an obvious way in which it can be used ideologically. In most cases, this is unconscious: the terms used in a thesaurus usually represent, in a relatively unfiltered way, the prevailing modes of expression at the time of its compilation. Because these modes have ideological underpinnings, in that they present a supposedly transparent reflection of the world, the metadata itself acquires ideological colourings. The world of descriptive metadata, particularly the use of terms to describe subjects, is inevitably tied up in ideological knots.

Examining some of the changes that have occurred to the Library of Congress' Subject Headings (LCSH), a widely-used list of subject terms created and administered by that august institution, reveals insights into changes in attitudes over the span of their existence. Some of these are cultural and societal that reflect changing perceptions of appropriate labelling. The current term *People with disabilities*, for instance, replaced its predecessor *Handicapped*, which in its turn had replaced the original heading *Cripples. Romanies* replaced *Gypsies*, *African Americans* replaced *Afro-Americans;* this last term replaced *Blacks* which had in turn superseded *Negroes*. What is often surprising is how late some of these changes were made: *Romanies* only appeared as a term in 2001 and *People with disabilities* in 2002 [4].

Also revealing are the use of terms that find their home within the spheres of religion, politics or sex. Religion can be a thorny area and examining changes to the LCSH can offer illuminating insights into the prevailing attitudes that surround it. Until 2006, the term *God* referred to the Christian God alone: it had to be qualified to refer to the gods of other religions (such as *God, Muslim*). For a pluralist country such as the United States (not to mention a pluralist world), this was certainly problematic. This was changed, again surprisingly late, so that the Christian God was no longer the default option for the deity [5].

Plenty of examples can also be taken from the area of politics. A notable one stems from the Vietnam War. Until 2006, there was no entry in LCSH under this heading: it was instead referred to as the *Vietnamese Conflict, 1961–1975*. The reason for this was that the United States had never officially declared war on North Vietnam and its involvement there was never considered by the US Government to have this status. Despite its numerous military operations overseas, the US has not officially declared war, which requires the approval of Congress, since the end of the Second World War. Describing the engagement in Vietnam as a conflict effectively aided this narrative. Only in 2006 did the Library of Congress take the logical step of changing the entry to match popular usage [4].

Sex and sexuality is also a subject where terminology can be a highly charged focus of metadata. A glance back to the classification scheme used by the Library of Congress in 1950 tells us that books on homosexuality were labelled under *Abnormal Sex Relations (including sexual crimes)*. Included in this category are *Homosexuality* itself*, Sadism, Masochism, Fetishism* etc., *Prostitution and Sexual Perversion in Woman* (there is no separate heading for sexual perversion in man). Only in 1971 was this changed when a new heading *Homosexuality, Lesbianism—Gay Liberation*

Movement, Homophile Movement saw the light of day. An apparently small change but one that was seen by one author at least as having an 'electrifying effect' on the gay rights movement of the time [6]. Terminology, even in libraries, is important.

Examples of this type can be found everywhere metadata is made and used: they are certainly not unique to libraries. It is as much a feature of language itself as of language in metadata. Just as revealing as the choice of words is where they fit into an overall scheme or classification. A taxonomy and its architecture can be as ideological as the terms with which it is populated.

Classification: Hierarchy and Ideology

As we saw in Chap. 2 and will explore in more detail in Chap. 6, one of the main ways in which humans have attempted to understand the world is to divide it into categories and arrange these into hierarchies. In libraries this way of organizing knowledge runs from the pioneering work of Kallimachos' *Pinakes* through to the Bodleian catalogues of the seventeenth century and to Dewey's renowned classification scheme.

What is expressed by the layers of a hierarchy is merely that its lower levels are part of those above them, that they contain narrower topics which find a place nested within the broader subjects one layer up. What exactly is meant by 'part of' or 'narrower' is often left vague if it is defined at all. It need not express any notion of superior or inferior, that upper levels are more important or significant than those below: in fact, most classification schemes tend to emphasize precision and so recommend using subjects that are as low down in the tree of subjects as possible.

Despite this, classification schemes may exhibit an ideological tinge when notions of relative rank are allowed to come into play; in these cases the levels in a hierarchy may act as tiers within which notions of superior and inferior can be embedded. These distinctions may also manifest themselves in the ordering of concepts within a single level. Often it is considered that the first topics listed on any layer have more weight or significance than those following; this can produce a secondary hierarchy nestled within the flat landscape of a single level.

A number of commonly-used classification schemes reveal something of an ideological agenda in their top-level subjects and the way in which they are ordered. One of the clearest is the Chinese Library Classification, the scheme used in almost all libraries and publishing operations in the People's Republic of China. The opening of the *Wikipedia* entry on the classification scheme shows the first 3 of its 26 top-level categories (labelled, logically, A to Z) (Fig. 4.3).

The first top-level category (A) in the scheme is *Marxism, Leninism & Deng Xiaoping Theory.* Immediately below this are the works of five authors evidently considered the most important theorists, Marx, Engels, Lenin, Stalin, and Mao: Deng Xiaoping, the Chinese leader credited with introducing market economics into China, also appears here but as a sub-category of Mao. Other top-level categories in this classification also reveal something of the priorities of those who

From Wikipedia, the free encyclopedia *Chinese Library Classification*

Chinese Library Classification

The **Chinese Library Classification** (?????? ?? CLC), also known as **Classification for Chinese Libraries** (CCL), is effectively the national library classification scheme in China. It is used in almost all primary and secondary schools, universities, academic institutions, as well as public libraries. It is also used by publishers to classify all books published in China.

The **Book Classification of Chinese Libraries** (BCCL) was first published in 1975, under the auspices of China's Administrative Bureau of Cultural Affairs. Its fourth edition (1999) was renamed CLC. CLC has twenty-two top-level categories, and inherits a Marxist orientation from its earlier editions[1]. (For instance, category A is Marxism, Leninism, Maoism & Deng Xiaoping Theory.) It contains a total of 43600 categories, many of which are recent additions, meeting the needs of a rapidly changing nation[2].

The CLC System

The 22 top categories and selected sub-categories of CLC (4th Edition) are as follows:

A. Marxism, Leninism, Maoism & Deng Xiaoping Theory

- A1 The Works of Karl Marx and Friedrich Engels
- A2 The Works of Vladimir Lenin
- A3 The Works of Joseph Stalin
- A4 The Works of Mao Zedong
  - A49 The works of Deng Xiaoping
- A5 The Symposium/Collection of Marx, Engels, Lenin, Stalin, Mao and Deng Xiaoping
- A7 The biobibliography and biography of Marx, Engels, Lenin, Stalin, Mao and Deng Xiaoping
- A8 Study and Research of Marxism, Leninism, Maoism & Deng Xiaoping Theory

B. Philosophy and Religion

- B0 Philosophical schools
- B1 Philosophy (Worldwide)
- B2 Philosophy in China
  - B22 Pre-Qin Dynasty Philosophy (~before 220 BC)
    - B222 The Confucian School
      - B222.2 Confucius (Kǒng Qiū, 551-479 BC)
- B3 Philosophy in Asia
- B4 Philosophy in Africa
- B5 Philosophy in Europe
- B6 Philosophy in Australasia
- B7 Philosophy in America
- B8 Cognitive science
- B9 Religion
  - B91 Sociology of Religion, Religion and Science
  - B92 Philosophy and History of Religion
  - B93 Mythology and Primitive religion
  - B94 Buddhism
  - B95 Taoism
  - B96 Islam
  - B97 Christianity
    - B971 Bible
      - B971.1 Old Testament
      - B971.2 New Testament
    - B972 Doctrine, Theology
    - B975 Evangelism, Sermon
    - B976 Christian Denomination
      - B976.1 Roman Catholic Church
      - B976.2 Orthodox Christianity (Eastern Orthodoxy, Oriental Orthodoxy)
      - B976.3 Protestantism (Protestant Reformation)
    - B977 Ecclesiastical polity
    - B978 Research on Christianity
    - B979 History of Christianity
      - B979.9 Biography
  - B98 Other Religions
  - B99 Augury, Superstition

C. Social Sciences

- C0 Social Scientific Theory and Methodology
- C1 Present and Future of Social Sciences
- C2 Organisations, Groups, Conferences
- C3 Method of Research in Social Sciences
- C4 Education and Popularization of Social Sciences
- C5 Serials, Anthologies, Periodicals in Social Sciences
- C6 Reference Materials in Social Sciences

Fig. 4.3 Chinese Library Classification opening categories (From Wikipedia)

compiled it. These include *Military Science* (E), *Culture, Science, Education and Sports* (G), *Agricultural Science* (S), *Industrial Technology* (T), *Transportation* (U) and *Aviation and Aerospace* (V), all areas in which the Chinese government has placed some emphasis since the People's Republic was founded in 1949.

Some evidence of an ideological agenda also appears in the ordering of topics within a level. After *Marxism, Leninism & Deng Xiaoping Theory* the top level features *Philosophy and Religion* (B), *Social Sciences* (C), and *Politics and Law* (D), all topics of concern in the context of a Marxist framework. Literature, art, history, geography and the sciences all appear much further down the list of categories. This is not so different from Dewey's ordering of top-level topics which we came across in Chap. 2: he also puts philosophy, religion and social science towards the beginning of his top layer, relegating the sciences, art, literature, history and geography towards the end.

All of which should be enough to make it clear that metadata as ideology is not confined to Marxism or other philosophies often labelled 'ideological' as a derogatory epithet. The classification scheme employed by the US Library of Congress also reveals something of the preoccupations of those who compiled it (Fig. 4.4).

Here we see that the history of the Americas (E & F) enjoys the same top-level status as the history of the rest of the world combined (D) and is considered so important that it merits two entries (one concerning the history of the United States alone (E), the other covering Canada, Mexico and Latin America lumped together (F)). Here we also see that the military and naval sciences receive recognition as separate top-level categories, an even more emphatic assertion of their importance than in the Chinese classification where they are combined into one.

There is nothing particularly sinister in this. Many classification schemes give priority to those topics that are likely to be of most interest to their community of users (after all, the primary purpose of the Library of Congress was initially to service members of Congress, who might well have interests in military and naval science). They are ideological only in the sense that Cook defines it, representing a view of the world as a transparent window into its supposed reality. But they are ideological nonetheless unless they are open about this.

More sinister uses of classification to serve a direct and often pernicious ideological purpose appear sporadically throughout history. It has often been used in some form to give supposedly 'scientific' credence to racist theories by dividing up humans into discrete racial groupings. Early examples of this include the work of the German anthropologist Johann Friedrich Blumenbach; in the late eighteenth century he came up with a five-fold division of races that was elaborated by such later proponents of his approach as Jean Baptiste Julien d'Omalius d'Halloy and Louis Figuier.

Even more insidious was the adoption of these approaches by eugenicists in the early twentieth century. One of most notorious of these was the American Lothrop Stoddard, the author of more than 20 racist diatribes in which he lamented the diminution of white supremacy by 'colored' races. Stoddard based much on his argument on his own racial classifications which not only divided humans by colour but also subdivided white races into 'Nordic', 'Alpine' and Mediterranean, of which he

Fig. 4.4 Library of Congress Classification: top-level concepts

considered Nordic the most superior. Stoddard's theories led him into eugenics and to encounters with the Nazis, whose bureaucracy for administering forced sterilization he gushingly praised in a travel memoir of a visit to Germany in 1940 [7].

The Nazis themselves used classification to provide a pseudo-scientific basis for their ideological purposes. The Nuremberg Laws of 1935, which limited citizenship to those considered ethnically German and forbade marriage between Germans and Jews, was backed up by a classification scheme to determine who exactly was Jewish and who was not. A chart from the period (Fig. 4.5) shows these racial classifications diagrammatically. Those with four white circles at the top left of the diagram were wholly German, these circles indicating their wholly German grandparents. Those with three or four black circles at the top of the two columns to the right (three or four Jewish grandparents) were classified as Jews, while the remaining two columns indicated those of mixed race.

Fig. 4.5 Chart of racial classifications under Nuremberg Laws (1935)

During the War, the Nazis also employed the supposedly scientific application of metadata to assign categories of Germaness to the inhabitants of occupied territories. The *Deutsche Volksliste*, devised under the auspices Heinrich Himmler, put these people into a sliding scale from *Volksdeutsche* (ethnically German) to *Rűckgedeutsche* (those considered worthy because of their racial background but who resisted turning into true Germans) [8]. The category to which a person was assigned could determine whether they ended up in the SS, the Wehrmacht or a concentration camp.

These examples are some of the most blatant uses of metadata to support ideological agendas. They are so effective because classification is such an important part of how humans understand the world: it sets the framework within which most of our conceptions of knowledge operate. These frameworks can be questioned occasionally but in general they are accepted as the way 'things are' (to quote Cook once again) and our intellectual endeavour then works to apply them. A chart such as that devised to explain the Nuremberg Laws serves its ideological purpose because the urge to taxonomy (the subject of Chap. 6) is such a potent one.

The Ideology of 'Objective' Metadata

There are plentiful examples of metadata as a tool of ideology, some of them blatant, most of them invisible as much to their creators as to those who are intended to use them. We could take the pedantic view of ideological metadata as a truism, reflecting a set of beliefs that consider themselves transparent reflections of reality. Or we could be more restrictive, and consider it ideology only when it is part of a clear agenda to promote these views as if they were reality.

There can be another ideological dimension to metadata beyond this. Can the way we think of metadata itself have ideological underpinnings? It can be such an effective ideological tool because it has a certain aura of objectivity to it, despite the fact that, as we saw in Chap. 1, it is nothing of the sort. The previous chapters should have made this abundantly clear for descriptive metadata, the type that is most obviously constructed by humans for humans. But what of the transactional, technical metadata which online systems gather to enable their operations – surely this is more objective?

This is the kind of metadata that was highlighted by Edward Snowden, what The Guardian newspaper described as "information generated as you use technology", the type he revealed the NSA to be collecting in huge quantities from our everyday online activities. Part of the rationale put forward to ease public concern on this was that because it is metadata, it is just an objective record of what is happening as we conduct our online transactions. This objectivity renders it less intrusive.

The truth, as the Dutch media academic José van Dijck points out, is far from this benign view. What is collected, how it is gathered and the form it takes are all determined by those who design and administer the systems that harvest it. What they choose to collect is far from a comprehensive and objective view of reality but is, she points out, "value-laden piles of code that are multivalent and should be approached as multi-interpretable data"[9].

But to maintain the fiction of objectivity and sugar the pill of intrusion that the gathering of this metadata entails requires an ideological underpinning, a belief in its impartiality and neutrality. Like all ideologies this must also be presented as a reflection of the way things are. Certainly the technical garb of this metadata and the way it is gathered, apparently free from human interference, makes it easier to present it in this way, but the ideology behind it, which claims that transactional metadata is an objective record, is just as important.

So metadata is steeped in ideology, as is almost any other human creation. It may have the veneer of something detached from its progenitors but it can never wholly cut its umbilical cord. But what exactly makes up metadata, what comprises the abstracted data that helps us change our information into knowledge? The next three chapters examine its make-up from its internal organs to the architectures within which the metadata organism resides.

References

1. Encyclopedia Britannica. (2015). Antoine-Louis-Claude, Comte Destutt de Tracy. http://www.britannica.com/biography/Antoine-Louis-Claude-Comte-Destutt-de-Tracy. Accessed 30 Oct 2015.
2. Williams, R. (1985). *Keywords: A vocabulary of culture and society*. Oxford: Oxford University Press.
3. Cook, N. (1998). *Music: A very short introduction*. Oxford: Oxford University Press.
4. Litwin, R. (2006). Library Juice » Interview with Barbara Tillett. *Library Juice*. http://libraryjuicepress.com/blog/?p=115. Accessed 10 Nov 2015.
5. Beall, J. (2006). Ethnic groups and Library of Congress Subject Headings. http://eprints.rclis.org/8831/1/ethnicgroups.pdf. Accessed 10 Nov 2015.
6. Kagan, A. (2015). *Progressive library organizations: A worldwide history*. Jefferson: McFarland.
7. Stoddard, L. (1940). *Into the darkness: Nazi Germany today*. New York: Duell, Sloan & Pearce.
8. Overy, R. J. (2004). *The dictators: Hitler's Germany and Stalin's Russia*. New York: W.W. Norton.
9. Van Dijck, J. (2014). Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology. *Surveillance and Society, 12*, 197–208.

Chapter 5
The Ontology of Metadata

The historical survey of metadata in Chaps. 2 and 3 ended with a picture of something of jungle of standards. They appear to be everywhere, proliferating wildly, constantly evolving and giving birth to offshoots which take on a life of their own. But what exactly is a metadata standard? As so often here, it is best to consult a librarian. Priscilla Caplan, in one of the most widely-read textbooks on metadata for library science students defines a standard as "a set of metadata elements and rules for their use which have been designed for a particular purpose" [1].

These are what is usually found in the often hundreds of pages that document each of these standards: a set of fields into which metadata can be slotted and a set of rules governing how they should be deployed and what should go into them. The contents of a standard inevitably reflect the make-up of metadata itself, its core components and the way in which they interact with each other. It is the ontology of metadata, what it is and what it is made of, that is the subject of this chapter.

Metadata is usually considered to have three fundamental components, some (but not necessarily all) of which will be defined in a standard. Metaphors from linguistics, somewhat loosely applied, are often used to define these. They are-

- *semantics:* the meanings of the fields or elements into which the metadata is put
- *syntax*: the way in which the metadata is encoded, perhaps in a spreadsheet, database table or a more generic format such as XML (eXtensible Markup Language) (of which there will be more later in this chapter).
- *content rules*: the rules, if any, which govern the content of the metadata itself, what is recorded, what form it should take and what should be excluded

Semantics

In linguistics semantics is the study of meaning, specifically the study of the relationship between a *signifier* (a symbol, perhaps a word, phrase or image) and its *signified* (what it refers to). As Saussure showed over a century ago, this relationship

R. Gartner, *Metadata*, DOI 10.1007/978-3-319-40893-4_5

Table 5.1 Definition of 'title' in five metadata standards

Metadata standard	Definition of title
Dublin Core	The name given to the resource. Typically, a title will be a name by which the resource is formally known
Anglo-American Cataloging Rules (AACR2)	The chief name of an item and includes any alternative title, but excludes parallel titles and other title information.
Encoded Archival Description	The name, either formal or supplied, of the described materials
VRA Core (visual objects)	The title or identifying phrase given to a work or an image
PBCore (public broadcasting)	A name or label relevant to the asset

is essentially arbitrary: there is no objective reason why the sequence of letters that spells *creator* should necessarily refer to someone who has brought something into existence; it acquires this meaning by the context in which it is found and its relationship to other sequences of letters [2].

In metadata, semantics is the generic term used to describe the relationship between the fields or elements of a standard and the content that fills them. A metadata standard will usually define a set of these fields (such as the 15 that make up Simple Dublin Core), explain what they mean and often indicate whether or not they must be present in a record. Dublin Core, for instance, specifies the name of its element *Coverage,* defines it as "the spatial or temporal topic of the resource, the spatial applicability of the resource, or the jurisdiction under which the resource is relevant.", and indicates that it is optional.

Saussure's point about the arbitrary nature of the sign applies to these fieldnames. There is no more reason why the spatial or temporal topic of a resource should be called *Coverage*, any more than it should be labelled *couverture*, *Umfang, kattavuus,* نطاق, 范围 or its equivalent in any other language. Nor does it make any more semantic sense to label it *Coverage* than 500, its equivalent in MARC. As for any sign, the label for the field *Coverage* acquires its meaning from its context, the fact that it is defined within Dublin Core standard. It takes on the meaning that the standard prescribes for it.

Knowing this context is vital to make sense of the semantics of a field. Even something as basic as a title can be defined in subtly different ways in different standards as is shown in Table 5.1.

These definitions vary noticeably in what they consider a title: they all see it as some type of 'name' for a resource but differ in what particular name should be accorded this status. AACR2 homes in specifically on the 'chief name', excluding such things as its equivalent in another language (known as a parallel title). VRA Core, a standard used to catalogue visual images, circuitously and rather unhelpfully defines a title as a 'title'. PBCore, a key standard used in the broadcasting industry, even more vaguely considers it to be a name 'relevant' to the resource, whatever 'relevant' means. So even if we know the standard in which a title is defined, we may still be somewhat confused as to exactly what is meant by it. This is perhaps where the linguistics metaphor should move away from semantics to pragmatics, the study of how meaning is acquired through the context of language in use.

Because it is so important to know the provenance of a field name in order to be clear about its semantics, the metadata world has found a more precise way of identifying these than a human-readable label such as *Title* or *Coverage*. This is done by using a string of letters, numbers and punctuation marks known as a *Uniform Resource Identifier* (usually abbreviated to URI). Instead of using the label *Creator* and indicating somewhere that we are talking about a *Creator* as defined in Dublin Core, we can use this URI to define the element precisely:-

```
http://purl.org/dc/elements/1.1/creator
```

The format of this string may look familiar to anyone who has used a browser such as Firefox to access the World Wide Web: it looks just like the address we put in to be taken to a website. In fact, a Web address, usually referred to as a *Uniform Resource Locator* (URL), is simply one example of a URI, in this case used to identify a place on the Web. URIs can have a much wider remit than this: they can identify *anything* on the Internet or even outside it, everything from abstract philosophical concepts to physical objects. They can also be used, as in the example above, to pin down the semantics of a metadata element.

The important point about a URI is that it should be unique anywhere on the Internet. This is made possible because of the first part of the identifier after the prefix *http://* In a Web address this would be the location of the website; in a URI, it is usually the authority that has defined it. These are unique across the Internet, which means that, even if every character in the URI after this initial string is replicated elsewhere, the URI itself is unique.

Using a URI instead of the label *Creator* means that we know that we are talking about a creator as defined by Dublin Core and not any other type. A URI need not be confined to the semantics of a metadata element in this way: it can also be used for the *content* that is put into the element and to express its relationships with others. As we shall see later, URIs form the backbone of the Semantic Web precisely because they pin down semantics uniquely and precisely.

Identifying clearly what the semantics of a metadata field are is one thing, reconciling the semantics of different schemes is another matter entirely. This is important because without this, it becomes difficult to share metadata or move it around with any certainty that it will be interpreted properly. 'Crosswalks' have been constructed between most major standards to allow this. These are mappings of field to field on the basis of similar semantics: there are, for example, crosswalks between Dublin Core and MARC, EAD and several other standards. Few can be entirely precise for the reasons shown in the *Title* example in Table 5.1: their definitions are rarely congruent, even if it is clear that they are trying to talk about the same thing.

Something that compounds this problem is the issue of semantic breadth. Some schemes use narrow, precise definitions for a single concept, often splitting it into several constituent fields, while others adopt a much broader approach, deliberately using wider definitions which are simpler to implement but less useful when precision is required. Dublin Core in particular deliberately defines broader semantics for its simple element set than its counterparts such as MARC. Moving data

from a narrower to a broader semantic element inevitably entails losing some detail or nuance. This loss may be justified if it allows metadata to be shared more widely (one of the rationales for a standard such as Dublin Core) but it can be difficult to prevent a dumbing-down of metadata which loses more than it gains.

Syntax

The second core component of metadata for which a linguistic metaphor is borrowed, rather more loosely than for semantics, is that of syntax. In linguistics syntax refers to the rules that govern how the components of language are linked together in structures to form units such as phrases or sentences. In metadata, it is used by analogy to describe the ways in which metadata is encoded, particularly in its machine-readable form. Metadata standards are usually designed to allow their contents to be interchangeable so that they can be transferred between systems. They employ syntax to enable these exchanges to become possible.

This need for this interchangeability, or *interoperability* to give it its more technical term, is both spatial and temporal. Metadata, like manure, is of limited value if not spread around: unlike manure, perhaps, we want it to last a long time. It will be of limited use to anyone if it is stuck in a proprietary piece of software which has formats that cannot be understood by any other package or are likely to become obsolete and unreadable before long. Here the analogy with language is useful: the syntactical structures of metadata, the rules that govern how we should encode its semantic components and their relationships to each other, are what allows it to do more than just talk to itself. It is what makes it able to communicate rather than merely record.

One format for encoding metadata (and a good deal of data) has achieved a degree of predominance: this goes by the unprepossessing name of *eXtensible Markup Language* or *XML*. XML started (under a different name and in a slightly more complex form) as a way of marking up electronic texts to allow more to be done with them than simply reading their contents on a computer. It was particularly adept at supporting the sophisticated use of texts in academic projects: large corpora of spoken languages could be compiled for linguistic analysis, for instance, or interactive scholarly editions of a medieval manuscript could readily be created that went well beyond what was possible in their traditional paper antecedents. Later its value as a medium for metadata became more widely recognized.

Finding an example of XML on the Internet is as easy as loading a website into a browser and choosing to view its page source. Figure 5.1 shows a page from Wikipedia and a small part of its (much simplified) XML coding that can be revealed by doing this. What appears here is the text of the web page itself marked up with a series of tags. These tags define *elements* within the XML file, its core components: they tell us, and the machines that process the file, something about the text that they surround. They can be nested within each other so that they encode their mutual relationships and introduce a structure to the text. Complex structures which go far

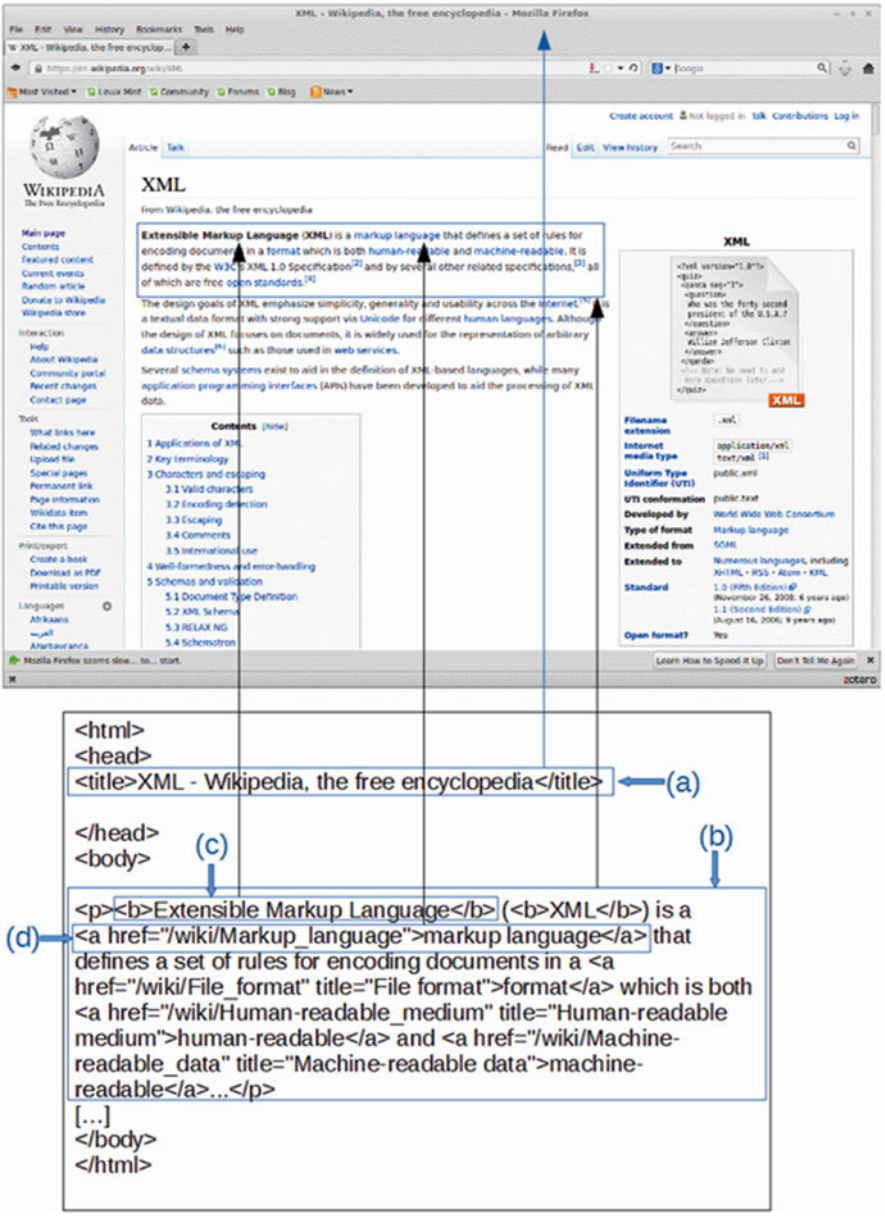


Fig. 5.1 A Wikipedia page and its underlying XML markup (highly simplified)

beyond the simple 'flat' files of a format such as a spreadsheet can soon be built up; these are often an essential part of a metadata standard.

Each element begins with its name in pointed brackets (as in the <title> element (marked **(a)** in the diagram) and ends with a tag which is identical except for the

addition of a forward-slash (/) before the element's name (</title>). In a web page they generally function as instructions to the browser for formatting the text they surround. In this example, the element <title> instructs the browser to print its contents (here the words "XML – Wikipedia, the free encyclopedia") on the bar at the top of the browser screen. The element <p>, **(b)**, tells the browser to treat the text it encloses as a separate paragraph, formatting it as such (usually with a blank line before and after). The element <b>, **(c)**, which stands for bold, simply tells the browser to embolden its contents.

Other elements are more complicated than these simple formatting instructions. The one marked **(d)** that reads:-

```
<a href="/wiki/Markup_language">markup language</a>
```

tells the browser that the words **markup language** are a link to another document, specifically one called *Markup_language*. This element, <a> (anchor), has more than just formatting instructions attached to it: it also includes the address of the other document. This information is given in what is known as an *attribute*, the part of the element's opening tag that here takes the form **href="/wiki/Markup_language "** . Attributes qualify an element by attaching additional semantics to its basic meaning. Here the browser knows not only to format this text in a given way but also how to react if the user clicks on it, in this case to open the document *Markup_language*.

This is a very simple use of XML which records instructions for formatting a text and behaving in particular ways when the user interacts with it. But XML is much more powerful as a syntax for encoding semantics, telling something about what the content of the tags mean rather than just how to process them. Take this example of a short text marked up in XML:-

```
<POEM>
   <TITLE>Lines Written in Early Spring</TITLE>
   <AUTHOR>
          <FIRSTNAME>William</FIRSTNAME>
           <LASTNAME>Wordsworth</LASTNAME>
    </AUTHOR>
   <STANZA>
         <LINE N="1">I heard a thousand blended notes,</LINE>
         <LINE N="2">While in grove I sate reclined,</LINE>
         <LINE N="3">In that sweet mood when pleasant thoughts</LINE>
         <LINE N="4">Bring sad thoughts to the mind.</LINE>
    </STANZA>
   <STANZA>
         <LINE N="5">To her fair works did nature link</LINE>
         <LINE N="6">The human soul that through me ran;</LINE>
         <LINE N="7">And much it griev'd me my heart to think</LINE>
         <LINE N="8">What man has made of man.</LINE>
    </STANZA>
</POEM>
```

This provides us with far more information than mere instructions on how it should be formatted in a web browser: we can infer that we are looking at a poem (the clue is the element <POEM>), that its title is *Lines Written in Early Spring* (from its <TITLE>), its author is William (his <FIRSTNAME>) Wordsworth (his <LASTNAME>) and it is divided into two <STANZA>s. We are recording more precise semantics here about the poem than in the case of the web page, although we are using the same language of tags, elements and attributes to do it.

What differs between these two examples is that we are in each case deploying a different set of elements and following different rules for how we apply them. A set of these elements and rules is known as an *XML schema.* There are hundreds of schemas around, each designed for a specific purpose. The most commonly used in the world today by a great margin is the one shown in the first example: it is known as HTML (Hypertext Markup Language), and is used to format every web page on the Internet. The other example uses an *ad hoc* schema written by the current author.

Both of these show XML encoding data, the text of a web page in the first, the words of a poem in the second. But it can be used just as easily for metadata. A short record in Simple Dublin Core could look something like this:-

```
  <metadata>
    <title>Utopia</title>
    <creator>Thomas More</creator>
    <description>Sir Thomas More's work, first
published in 1516, describes an island community
free of private property and religious
intolerance</description>
    <publisher>Penguin Books</publisher>
    <date>1965</date>
  </metadata>
```

Many standards are issued as XML schemas alone and metadata that conforms to them must be encoded in this format. EAD, the scheme for archival descriptions, is one example of a standard that insists on this syntax. Others allow XML as an option and publish schemas to enable this but do not require that these are used if other formats are preferred: PREMIS (PREservation Metadata: Implementation Strategies) [3], a commonly-used standard for digital preservation, is an example of this. Dublin Core, perhaps the *lingua franca* of metadata, is published as a list of elements rather than an XML schema, but, as can be seen in the example above, is readily expressed in this format.

The reasons why XML is so popular for encoding metadata are several and compelling. It is not tied to any given software package and so will not become obsolete when an application crosses the digital Styx into oblivion. It is easy to move around between systems as it is encoded as text, just about the simplest format available. It is probably the most robust around for archiving data because of this simplicity. It is good at encoding hierarchies, one of the preferred ways for organizing metadata in many a standard. For all of these reasons, it is the closest that metadata has come to a common syntax.

Content Rules

The third component of metadata governs not the range of elements defined in a standard nor the way they are encoded but what goes into them, the content with which they are populated when the standard is in use: these are known, rather prosaically, as *content rules*.

We can usually get a rough idea of what should go into an element from its definition: the one for the Dublin Core *Title* element, for instance, tells us that the "name given to the resource" should go there. But this is still pretty vague: we usually need much more detail than this. We might want to know exactly what title should be put here (the one on a title page, perhaps, or one given by a cataloguer when there isn't a title page, the one in the original language of a text or its translation?). We would probably also need to be told what form it should take (should it be transliterated if it is in a foreign script, for example?). For this more detailed signposting we need content rules.

The key rationale behind these rules is consistency: only by applying them can metadata be found in a predictable and reliable way. Universal consistency, akin to reverse engineering a Tower of Babel, is never going to be achieved, but by at least enforcing some degree of regulation within a particular area (such as libraries) we can begin to feel confident that our metadata is as findable as it reasonably can be. That sounds like a simple ambition but drawing up a set of rules to cope with every contingency would require omniscience beyond the remit of any mere mortal.

Rules can often be formulated as answers to specific questions. If a journal article has 500 authors (common in the sciences) should they all be listed or only a select few? Should a person be listed under their real name or their pseudonym if that is how they are better known? How do we cope with compound names such as Laurens van der Post or George Bernard Shaw? Should we call a Lord by his title of ennoblement or his original name?

Other issues that can benefit from content rules are a little more perplexing. Take the case of a woman called Rosemary Brown, a middle-aged widow from London who achieved some fame in the 1960s and 1970s. Mrs Brown was a clairvoyant who claimed that the spirits of dead composers, including in their august company Beethoven, Schubert and Chopin, dictated their posthumous works to her. These she duly transcribed and released to the world: some of these were even recorded and issued on CD. If a disc of these works were to turn up in a library, should it be catalogued under 'Rosemary Brown' or the names of the composers who supposedly dictated them? An ethical librarian (for they are most certainly highly ethical people) would never dream of making a value judgment about the validity of her claims and so some rule is needed to allow for this unlikely contingency.

All of these vexing issues have been duly considered and rules written by bodies of experts to address them. The library world has, as usual in the area of metadata, been one of the more methodical in drawing up content rules: the most extensive of these are the venerable *Anglo-American Cataloguing Rules, 2nd edition* (usually abbreviated to AACR2) [4] and its intended, though not yet fully adopted, successor *Resource Description and Access* (RDA) [5].

AACR2, a hefty volume of rules that first appeared (as plain AACR) in 1967, remains the most-widely adopted cataloguing convention in the library world. Within its covers we can find answers to the perplexing questions posed above:-

- if a publication has more four or more authors, don't attempt to list them all, just list it under its title (rule 21.6C2)
- authors should be listed under their most commonly-used name: if this is a pseudonym, such as George Eliot, this is what should be used (rule 22.1)
- Laurens van der Post should be listed as "Van der Post, Laurens" (rule 22.5D) but George Bernard Shaw under "Shaw, Bernard" (rule 22.3A)
- Lords are listed under their title of nobility if that is the form by which they are most commonly known (rule 22.6)
- and finally, the vexed issue of the ghosts of long-dead composers: spirit communications should be listed under the name of the spirit and so Rosemary Brown's transcriptions would be listed under "Chopin, Frederic (Spirit)" and so on (rule 21.26)

This set of rules, extensive though it is, cannot cover every contingency, even within the narrow remit for which it was compiled, supplying conventions for the cataloguing of works in library collections. Over the Internet as a whole applying rules is almost impossible, if only because there is no central authority to enforce them. This does not mean that they are not applied in areas such as the digital media or digital commerce: on the contrary, it is clear that a site such as *Amazon* follows consistent metadata conventions to allow its complex and rapidly-changing operation to function. But these tend to be specific to a particular service: there is little sense yet that there is emerging a universal set of rules comparable to AACR2's place in the library world.

Controlled Vocabularies

Another way of enforcing some sense of consistency in the content of metadata is to limit it to restricted sets of allowed terms. Instead of issuing a complex set of rules for the cataloguer to follow, we instead provide them with a list of the names, subjects and other particulars that they are allowed to use to populate the metadata record. The usual term for these lists is *controlled vocabularies*.

One area that can clearly benefit from this is place names. Take the Belgian city of Antwerp. In English, it goes by this name although its native Flemish speakers call it Antwerpen: its French speakers refer to it as Anvers. It is addressed by at least another 28 variant names when referred to in works of literature and history; some of these, such as Ambivaritum, Anyeŕsah and Handoverpia, would at first glance be hard to recognize as referring to the same city. Although all of these variants are easily incorporated into a modern information retrieval system, it is still useful to use a preferred version of the name to cluster them together and make it clear that they all refer to the same place.

The problem is even more acute when it comes to people. This is particularly important when many share the same name. There are hundreds of John Smiths listed in the British Library catalogue: sorting out which is which is essential if we are looking for one of their works. It would certainly be useful to distinguish the British entertainer Bruce Forsyth from the Bruce Forsyth who edited a book on the *Position-sensitive detection of thermal neutrons* [6] or the Phil Collins of the pop group Genesis from the author of a book on using systems theory to lead church congregations [7]. Getting those wrong could cause some embarrassing confusion at the very least.

Long lists of names have been put together to help resolve these dilemmas; the only problem here is that, like metadata standards, there are so many to choose from. Geographic names receive the controlled vocabulary treatment in a compendious thesaurus compiled by the John Paul Getty Foundation [8]: there are over one million of them here, recorded with any variant forms, including their historical versions. There are plenty of rivals to the Getty Foundation's *magnum opus*, including lists for most countries in the world and even for extraterrestrial bodies [9].

Personal names are also the subject of many an attempt to list them and their variant forms. The Library of Congress has produced one of the longest-established of these, the *Library of Congress Name Authority File (LCNAF)* [10], a mammoth compilation of over eight million. Each has a URI, that Internet-wide identifier: even someone as obscure as the author of this volume has received one of these, http://lccn.loc.gov/nb99003434, which can be used to identify him wherever he lurks in the digital undergrowth.

Because of its origin in the library world, the LCNAF tends to list people who have either written something that appears in the Library's collections or are mentioned in them. A more recent service aims to list those whose contributions are more widely spread, including anyone associated with producing or distributing a creative work (broadly defined). The *International Standard Name Identifier (ISNI)* [11] lists more than eight million of these, including, as it says on its website, "researchers, inventors, writers, artists, visual creators, performers, producers, publishers, aggregators" and many others. Each gets a 16-digit code, akin to the 13 digits that books receive in the form of an ISBN. If we come across a Bruce Forsyth with the ISNI number **0000 0003 6150 1602**, we can sleep more easily knowing that we're talking about the host of *Strictly Come Dancing* and not his physicist namesake.

Controlling names, whatever they refer to, is a vital part of making metadata usable on a large scale, but just as important is controlling subjects, making them consistent enough to ensure that searching by topic can be a reasonably precise operation. This is a much more complex area, if only because asserting what an object is 'about' is to make much more of a value-judgment than just recording its name.

Although it may be feasible to deal with subjects by putting them into lengthy alphabetical lists as if they were names, this rapidly becomes difficult to implement when they grow to any significant size. For this reason, subject lists are usually arranged by grouping together related concepts and specifying the relationships

between them. For centuries until the present day, these have usually be arranged in hierarchies. It is this urge to taxonomy, to classify and arrange hierarchically, that is the subject of the next chapter.

References

1. Caplan, P. (2003). *Metadata fundamentals for all librarians*. Chicago: American Library Association.
2. de Saussure, F. (1993). *Course in general linguistics*. London: Duckworth.
3. Library of Congress. (2011). *PREMIS: Preservation metadata maintenance activity* (Library of Congress). http://www.loc.gov/standards/premis/. Accessed 28 Jan 2010.
4. American Library Association. (2012). Anglo-American Cataloging Rules homepage. http://www.aacr2.org/. Accessed 28 June 2013.
5. Library of Congress. (2013). *Resource Description and Access (RDA): Information and resources in preparation for RDA*. http://www.loc.gov/aba/rda/. Accessed 28 June 2013.
6. Convert, P., & Forsyth, J. B. (1983). *Position-sensitive detection of thermal neutrons*. London: Academic.
7. Stevens, R. P., & Collins, P. (1993). *The equipping pastor: A systems approach to congregational leadership*. London: Rowman & Littlefield.
8. Getty Research Institute. (2015). *Getty vocabularies* (Getty Research Institute). http://www.getty.edu/research/tools/vocabularies/index.html. Accessed 24 Sept 2015.
9. International Astronomical Union. (2015). Planetary Names: Welcome. http://planetarynames.wr.usgs.gov/. Accessed 24 Sept 2015.
10. Library of Congress. (2015). Library of Congress Authorities. http://authorities.loc.gov/. Accessed 31 Aug 2010.
11. ISNI. (2013). International Standard Name Identifier (ISNI). http://www.isni.org/. Accessed 3 June 2013.

Chapter 6
The Taxonomic Urge

Saying what something is "about", what its subject is, is where metadata is most nakedly a human creation. This is where it is most clearly an interpretation of the world, a filtering of its mass of data and information to create a digest, often a single word, of what it means to us. There have usually been two stages to making these assertions of "aboutness". The first consists of the relatively straightforward task of choosing a label to describe it. The next, rather more complex, step is to put these labels and the concepts they refer to into some type of structure in which they take on additional meaning by their positions and their relationships to each other. The creation and shaping of these structures is usually known as *taxonomy*, the 'science' (in a relatively loose sense) of classification.

Back to etymology: the term taxonomy derives from two Greek words, τάξις (an ordering or positioning) and νόμος (a law or principle). By asserting in its name that it is following laws or principles, it is an idea which extends beyond merely putting concepts or things into categories but also encompasses the overarching rules governing these categories and the ways in which they relate to each other. In this way it attempts to take on the mantle of a scientific discipline, and it is in the sciences, most notably biology, that the idea of taxonomy has its firmest roots.

How fundamental the notion of taxonomy is to human beings is a matter of some contention, but many major figures from the discipline of anthropology have emphasized how deeply embedded in human culture is the need to put the world into categories. One of the most influential anthropological works of the last century, Mary Douglas' *Purity and Danger*, claims convincingly that objects considered dirty or repulsive in many societies are seen in this way because they do not fit cleanly into obvious groupings. She analyzes the Abominations of Leviticus, asserting that the reason why pigs are not considered kosher is because they fall ambiguously into two categories, the cloven-hoofed animal and those that chew the cud. She famously describes dirt as 'matter out of place' [1], showing that it is not just boundary-crossing but also being sited in the wrong location, physical or conceptual, that can induce revulsion.

R. Gartner, *Metadata*, DOI 10.1007/978-3-319-40893-4_6

Although Douglas retracted her views on the origins of kosher dietary rules in a later edition of the work, her ideas on the almost visceral need to categorize remain convincing. She talks less of how fundamental it may be for us to put these categories into the structures that we understand as a taxonomy. For an insight into this, we can go back 50 years before her groundbreaking work to that of her forebear, the great French sociologist Emile Durkheim. He pointed out as early as 1912 that these structures may underlie key features of religious thought. He claims in his *The Elementary Forms of Religious Life* that taxonomy is a fundamental way in which religion is linked to society as a whole and specifically to social groupings within it. In particular, he asserts that the religious practice of totemism, the notion of a spiritual connection between humans and physical objects (including animals), is a way of linking social classifications (such as clans) to conceptions of cosmology [2].

Durkheim points out that the classification schemes underlying religion are *hierarchical*. There are, he says, "dominating members and others which are subordinate to the first…the different species of a single class are conceived as all placed on the same level in regard to each other" [2]. In pointing this out, he highlights a core feature of most taxonomies: they are arranged in layers, often with some notion that the higher levels contain fewer members but are in some way superior to those nested beneath them. The urge to classify also appears to be an urge to discriminate into superior and inferior.

In practice, hierarchies are not essential to taxonomy and some taxonomies operate without them. Nor need the levels of a hierarchy necessarily imply notions of superiority or inferiority. But it is such a common feature of the way humans classify the world that we might consider hierarchy a fundamental feature of cognition and even language. We could certainly look at Noam Chomsky's concept of the structure of a simple sentence in his hugely influential *Syntactic Structures* [3] (Fig. 6.1) to see how deeply embedded they appear to be in our linguistic makeup.

If hierarchies are so pervasive in human thinking, it is no surprise to find them figuring conspicuously in the ways in which we classify the world This is particularly so in the biological sciences where the pioneering work of the eighteenth-century Swedish botanist and zoologist Carl Linnaeus stands out as one of the great human endeavours in taxonomy. It was Linnaeus who first proposed the hierarchical arrangement of nature (*kingdom–order–family-genus-species*) which is still used to classify all living beings: to this day, any organism is known by the combination of its genus and species, such as *Homo sapiens*, according to the Linnaean scheme.

Taxonomy as Metadata: The World in Hierarchies

Unsurprisingly given their seeming ubiquity, hierarchies are conspicuous figures in the history of metadata. We saw in Chap. 2 that some of the earliest attempts at classification were hierarchical: Kallimachos's *Pinakes*, the Bodleian Library's 1620 catalogue and the Dewey Decimal Classification all made use of increasingly complex and sophisticated structures of this kind. In the world of archives, they have

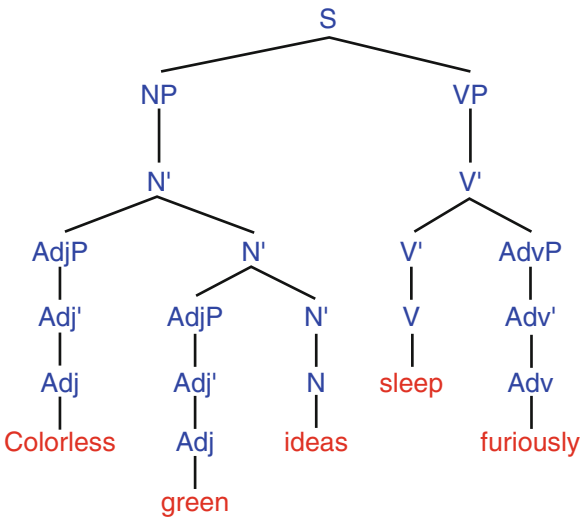


Fig. 6.1 The hierarchical syntactic structure of a simple sentence according to Chomsky (Diagram by Aaron Rotenberg)

been used in finding aids to organize collections, a practice that continues to this day in the *Encoded Archival Description (EAD).*

The Dewey Decimal Classification is a prime example of a taxonomic hierarchy on a grand scale. Although it may not be immediately obvious when browsing the shelves, the Dewey number is in fact a chain of subjects, from broad to narrow, all strung together in an order defined by Dewey and those who revised his *magnum opus* in later years. One way to see his scheme is as a flat-topped pyramid, akin to the famous stepped pyramid at Saqqara in Egypt. At the top lie the ten divisions of knowledge that we saw in Chap. 2, broad topics such as religion, social sciences, language and literature. Each of these is divided into successively more precise and narrower sub-topics. Every subject in the pyramid is given a number reflecting its journey down to its allotted location. The subject of oceanography, for instance, receives the number 551.46 because it has a precise slot of this kind (Fig. 6.2).

Each digit in the number represents one step down the pyramid into greater detail and greater precision. Because the scheme is based on our base-ten counting system, each of these steps is limited to ten subdivisions only; this introduces some potential distortions into an ideal partition of knowledge in cases where more than ten would be useful. But the use of numbers after the decimal point at least ensures that there need be no truncation of increasing precision when this is required. The longest Dewey Classification number currently recognized runs to a huge 27 digits, **331.892829225209712743090511** [4]; if Dewey's scheme can accommodate the Canadian tractor industry, the subject of this lengthy string, it should be able to cope with almost anything.

The technical term for this type of classification, in which each item is allotted a single place in a rigid hierarchy, is an *enumerative* scheme. In essence, a scheme is enumerative if every potential term that we might want to use is listed somewhere within its hierarchies and is given a single, immutable place within them. The

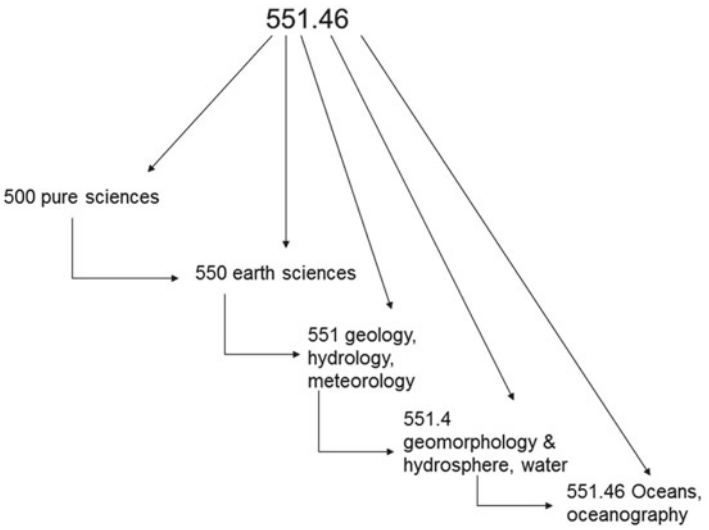


Fig. 6.2 Oceanography in the Dewey Decimal Classification

Dewey number finds a precise slot for a subject by chaining together a series of topics, from broad to specific, that lie above it in the pyramid. The concepts that are combined and the order in which they are presented are determined by Dewey and his successors. The combination of concepts is in no way left to the person doing a search.

This way of combining simple subjects to create more complex ones is not just for those few who construct classification schemes. Many a library cataloguer makes use of it to describe the complex subjects of works in their collections. A book on the great Indian librarian S.R. Ranganathan, whom we met in Chap. 2, might have a subject heading such as

```
Librarians--India—Biography
```

in which three very disparate concepts are joined together. In creating this chain, the cataloguer must second-guess what the person who uses the record will be looking for: they are assembling complex subjects which they judge will have some meaning to the searcher.

This method is usually given the technical name *pre-coordinate indexing*. One succinct definition comes from the Society of American Archivists who characterize it as:-

> A method of indexing materials that combines separate concepts in an item into a single heading [5]

One of the most compelling reasons for using the pre-coordinate approach is that it allows us to place a specific subject within its wider context. We saw in Chap. 1

that we can look on knowledge as being created by the cumulative forging of semantic links, starting with those between data and information. Pre-coordination puts a small semantic unit, a single subject, into a wider context (potentially the entire hierarchy of a classification). It could be argued that it is an expressive way of codifying ladders of concepts and so making the move from information to knowledge easier to achieve.

But is could equally be argued that it is an ambitious and potentially arrogant assumption that those who create compound subjects in this way have the capacity to define their contexts and the connections between the components that define them. It could be argued that pre-coordination, by erecting these mighty hierarchical edifices, moves away from building knowledge to ossifying it.

Thesauri: Introducing Flexibility into Hierarchies

Another approach to building a classification allows for a little more flexibility than the monumental hierarchies of a strictly enumerative scheme. To see how this is done, we need to go back over 250 years to the first publication of a famous reference work by the British lexicographer Peter Mark Roget. Roget had toyed, he himself claimed, with the idea of compiling a "system of verbal classification" for over 50 years before he published his life's work, the "classed catalogue of words" that still bears the title *Roget's Thesaurus* [6].

The name of Roget's *magnum opus* derives from the ancient Greek word for treasure house, a rather grandiose but apt view, from the perspective of a lexicographer at least, for an attempt to divide up the world of knowledge on the basis of the words used to describe it. First published in 1852, Roget's work tries to partition knowledge into a large hierarchy the shape of which owes its philosophical origins to such great names as Aristotle and Leibniz. At the summit lie six broad classes: 'abstract relations', 'space', 'matter', 'intellectual faculties', 'voluntary powers' and 'sentient and moral powers'. Over a thousand branches of sub-classes flow from these, building a tree of increasing density until we reach the 15,000 terms (a larger number in later editions) at the bottom of this mighty structure. Roget's work is still in print and is a much valued reference tool by writers searching for the *mot juste* to express precisely the ideas they seek to communicate.

In the late 1950s the world of information appropriated Roget's title when it was suggested that some means was needed to harmonize the language being used to index documents in order to make their retrieval more efficient [7]. By analogy with his model of a catalogue of words, the term *thesaurus* came to mean a new way of defining controlled vocabularies in a particular sphere of knowledge; this new approach remained hierarchical but more flexible than a rigid enumerative scheme.

In a modern thesaurus, certain terms are marked as 'preferred': these are the ones that should be used when classifying a document. They are put into context by being linked to broader terms higher up the hierarchy, narrower terms lower down and related terms on the same level. This is best illustrated by looking at an example of

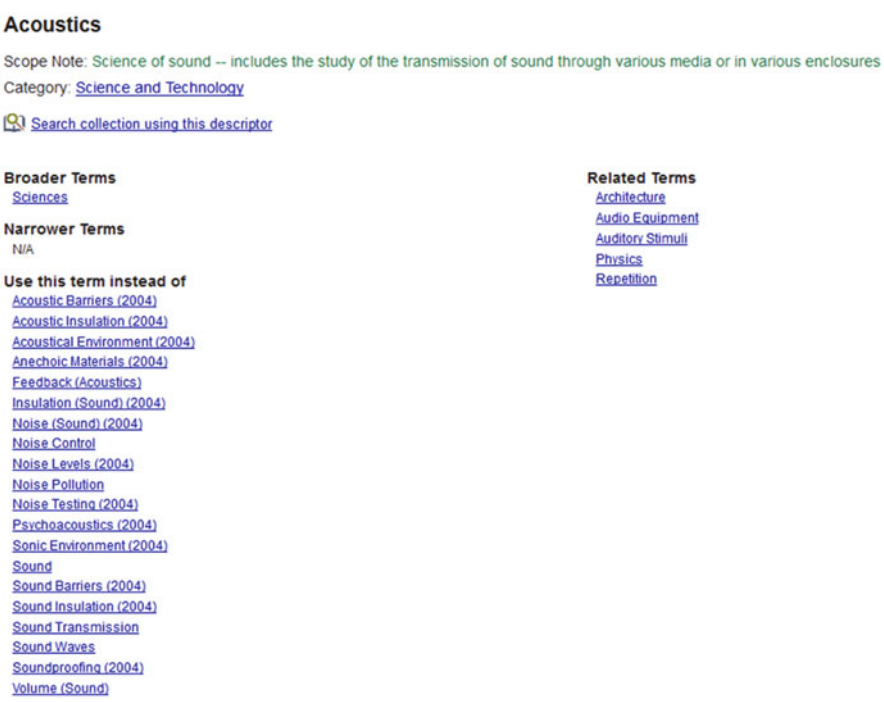


Fig. 6.3 Entry on Acoustics from ERIC thesaurus (Online version)

a working thesaurus. Figure 6.3 shows a sample entry from the well-known ERIC (Education Resources Information Center) thesaurus of terms for indexing literature on education research [8].

Highlighted at the top is the preferred term, *Acoustics*, the heading that the thesaurus recommends using for this subject: below this is a scope note, a short description of its meaning. Underneath these is a broader term *Sciences,* located higher up the ERIC hierarchy, and a number of related terms which are on the same level and can be used if more relevant than the preferred term itself. The thesaurus also includes others which are not recommended for use including *Sound, Sound Transmission and Sound Waves:* it suggests that *Acoustics* should be used instead of these. There are no narrower terms for acoustics, and so this term is the most precise one that covers this concept in the thesaurus.

What we have here is a large hierarchical taxonomy dissected and rearranged alphabetically so that each term forms a separate entry and its place in the hierarchy is shown by pointers to its broader, narrower and related counterparts. The thesaurus itself follows a strict hierarchical arrangement but this is an organizing principle that is often opaque to its users. Some search systems may offer full access to the thesaurus to allow users to find their way through its levels to the subject they want, but often all they see are the index terms chosen by the cataloguer who has followed this process in compiling the record.

Although a thesaurus is an expression of a hierarchy, it is a more flexible one than an enumerative scheme such as Dewey's. It is perfectly feasible for a subject term to find more than one place in its structure: Mary Douglas' repulsive pig, for instance, could readily fit into the categories of both 'cloven-hoofed' and 'cud-chewing' animals without causing any problems. This is simply done by introducing multiple broader terms for a subject: these 'polyhierarchies' can be valuable in representing the complexities of a subject without shoehorning a term into a single ill-fitting slot.

One drawback to the standard thesaurus is that the words 'broader', 'narrower' and 'related' do not tell us a great deal about the semantic relationships between terms: they are themselves rather fuzzy concepts. Usually, as a well-known article on thesaurus design points out, the broader-narrower relation indicates that an entry is either a type of another category (a cow is a type of mammal), a part of it (a finger is part of the hand) or an instance of it (Halley's comet is an instance of a comet) [9]. But we have to infer which of these applies from our reading of the thesaurus entry, a way of doing things that is imprecise to say the least. As we shall see later, information science has more recently come up with more explicit ways of describing semantic linkages in the form of the Semantic Web.

One great advantage that thesauri offer over their uncompromising enumerative forebears is that they make it possible to change the ways in which complex subjects can be handled when metadata is digital. When searching a computerized database it no longer makes sense for a complex subject to be defined in advance by a cataloguer: these technologies allow users to combine the terms they want, as they are searching, to define exactly the topics that interest *them*. To find the book on Ranganathan mentioned earlier it is no longer necessary to use the exact term defined by the cataloguer:-

```
Librarians--India—Biography
```

but instead as few or as many of these terms as are needed to express the concept that the searcher wants can be used instead. We could find the book with just two:-

```
Librarians AND Biography
```

if our interests were in the biography of librarians worldwide: we would miss it using these terms alone if the catalogue we searched insisted on the pre-coordinate subject entry, with its three terms, as the only way to find this book.

This approach, in which compound subjects are created by the user when they are conducting their search, has the technical term *post-coordinate* searching. One common way in which it operates is to allow searchers to combine terms as in the example above, linking them with the word AND to find only those matches that contain both. This method, known as Boolean searching, remains a very common way of finding precise matches, particularly in library catalogues or research datasets. But post-coordinate searching also operates whenever we put a term of more

than one word into Google – we are doing such a search here although the ways in which this site deals with the combination of words we enter is much more sophisticated than a simple Boolean query.

A Flowering of Taxonomies

Post-coordinate searching is now everywhere but this does not mean that hierarchies are dead. They may no longer be the one and only true path to find a subject but they remain one of the most common ways of organizing our view of the world. They now tend to act as a guide, an aid to navigation of our knowledge, rather than a prescribed route to it.

As taxonomy owes much of its origins to the work of Linnaeus, it is no surprise to find it flourishing in the area of biology. Here hierarchical models still predominate, particularly in evolutionary biology, the branch that concerns itself with mapping out the shared ancestry of organisms. Representations such as these look very much like modern descendants of the *Tree of Life*, a favourite model of Charles Darwin who speaks of it as an apposite metaphor for "the affinities of all beings of the same class" [10].

We find the same metaphor behind modern taxonomies which map evolution across species, although sometimes the plant in question is more succulent than arboreal. One very common representation is the cladogram (Fig. 6.4), a hierarchical structure which classifies organisms by their shared characteristics. Another is the Bayes cactus (Fig. 6.5), another representation of evolution and its effects, this time representing new levels as smaller buds stemming from their larger parents.

Hierarchical taxonomies are also common in the medical sciences. The classification scheme for diseases compiled and maintained by the World Health Organisation, the *International Statistical Classification of Diseases and Related Health Problems (ICD)* [11] is a prime example here. This mighty taxonomy contains over 14,000 codes covering diseases, symptoms and causes: it is an indispensable diagnostic tool which relies heavily on hierarchies for navigating the huge mass of information it contains.

Business has also embraced taxonomy with gusto. Many commercial enterprises now invest in knowledge management systems, ways of retaining the know-how and corporate wisdom (as they see it) that accumulates as they pursue their endeavours. These are often hugely complex systems which require careful organization of the knowledge that is to be preserved, shared and re-used. Complex classification schemes, usually referred to by the generic term *corporate taxonomy*, often underlie the information architectures of these. Unlike the ossified structures of enumerative schemes, these have to evolve rapidly and deal with large volumes of rapidly-growing information; they are much more flexible although most still have hierarchies at their centre.

We can also find the stratified model of hierarchy in at least part of the architecture of some well-known e-commerce sites. eBay, the popular online auction site,

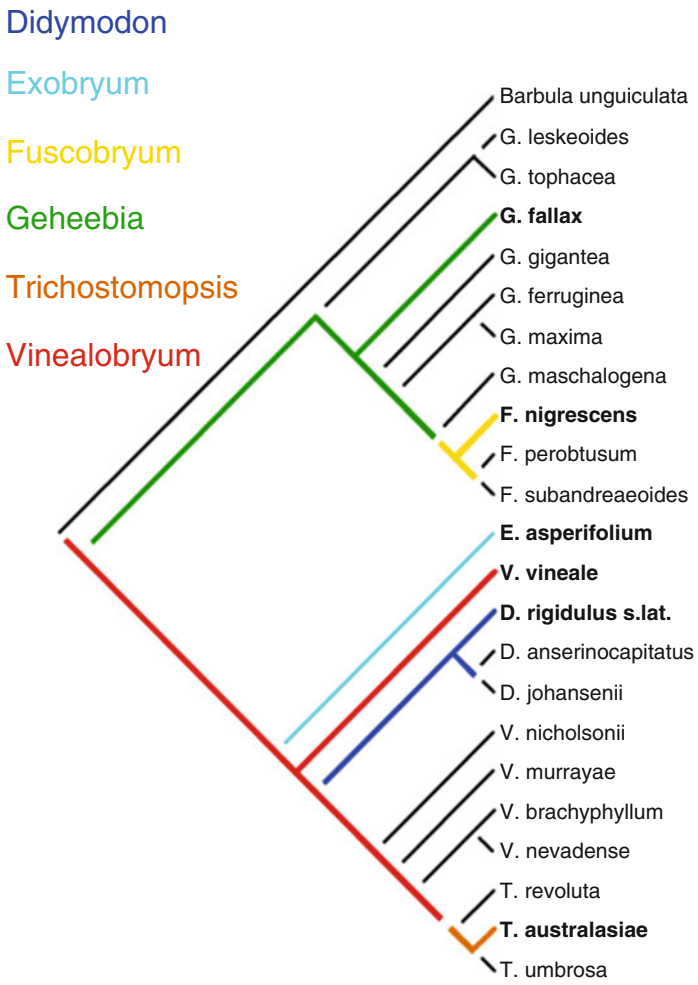


Fig. 6.4 A cladogram grouping organisms by shared characteristics in a tree-like structure (Diagram by Richard H. Zander)

organizes several thousand categories in this way to enable buyers to find what they want with minimal effort. When listing an item for sale, the seller is taken through the steps of a hierarchy to its most relevant slot and here is where the sale is listed. This tried-and-tested method appears again and again in online stores.

Hierarchies are clearly far from dead. They appear to have left an indelible imprint on human thinking and the metadata we employ to make the fruits of our intellectual labours usable and manageable. Ulysses' words from Shakespeare's *Troilus and Cressida* perhaps say something of the comfort blanket they offer:-

> Take but degree away, untune that string,
> And, hark, what discord follows! (Act 1, Scene 3).

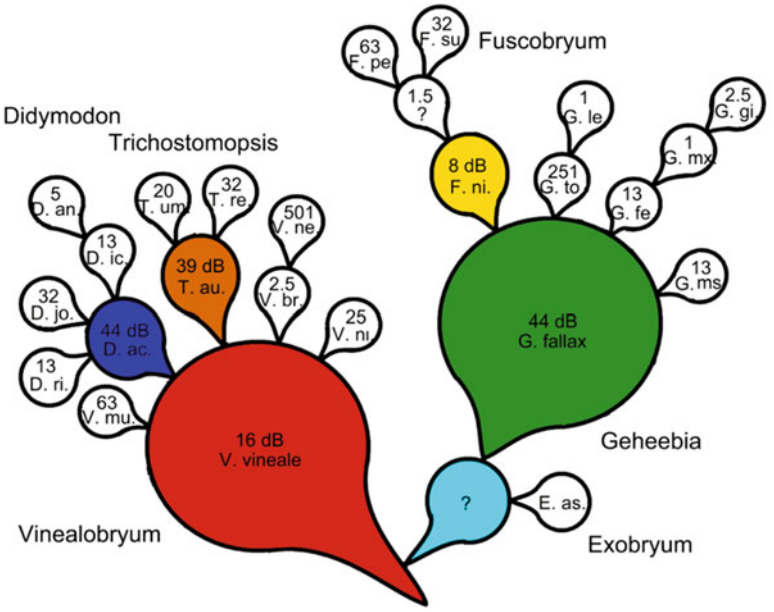


Fig. 6.5 A Bayes cactus, another representation of hierarchy (Diagram by Richard H. Zander)

The urge to taxonomy may be too strong to relinquish. But despite the hold it seems to have on us, other ways of organizing information and transforming it into knowledge have long been mooted. These have centred on moving away from hierarchies towards networks of interconnected information. It is this shift that forms the subject of the next chapter.

References

1. Douglas, M. (1966). *Purity and danger: An analysis of concepts of pollution and taboo*. London: Routledge and Kegan Paul.
2. Durkheim, E. (1915). *The elementary forms of the religious life*. London: Allen and Unwin.
3. Chomsky, N. (1957). *Syntactic structures*. The Hague: Mouton and Company.
4. Dewey Blog. (2006). 025.431: The Dewey blog: Exciting tractor-related news. http://ddc.typepad.com/025431/2006/02/exciting_tracto.html. Accessed 5 Oct 2015.
5. Society of American Archivists. (2015). Precoordinate indexing | Society of American Archivists. http://www2.archivists.org/glossary/terms/p/precoordinate-indexing. Accessed 5 Oct 2015.
6. Roget, P. M., & Kirkpatrick, B. (1987). *Roget's thesaurus of English words and phrases*. London: Longman.
7. Prasher, R. G., & Mangla, P. B. (1997). *Library and information science: Information science, information technology and its application*. New Delhi: Concept Publishing Company.
8. Educational Resources Information Center (U.S.), & Houston, J. E. (2001). *Vocabulary links: Thesaurus of ERIC descriptors*. Phoenix: Oryx Press.

9. Weinberg, B. H. (1998). Thesaurus design for information systems. http://www.allegrotechindexing.com/article02.htm. Accessed 29 Sept 2015.
10. Darwin, C. (1860). *On the origin of species by means of natural selection, or, the preservation of favoured races in the struggle for life*. London: J. Murray.
11. World Health Organisation. (2015). International Classification of Diseases (ICD). http://www.who.int/classifications/icd/en/. Accessed 13 Oct 2015.

Chapter 7
From Hierarchies to Networks

The ubiquity of hierarchies in metadata is by no means absolute, pervasive though they are. The problem with them has always been that they are inflexible and often ossify thought as much as they foster it. They can build statues which set our knowledge in stone instead of making it malleable and able to grow. The anthropologist Mary Douglas, whose comments on the importance of taxonomic boundaries we met in the last chapter, has been quoted in a pithy (albeit unsourced) comment as pointing out that "hierarchy works well in a stable environment" [1]. This is undoubtedly true, but little in human knowledge has been stable enough to make hierarchy necessarily the optimal way of dealing with it.

No more so has this been true as in the last few years now that the Internet and the revolution it has brought about in the transfer of knowledge have made their full impact on the world. The extent of the problem was highlighted as long ago as 1981 by the renowned architect, inventor and polymath Buckminster Fuller. He famously demonstrated how quickly the speed with which knowledge accumulates has accelerated over human history in his *Knowledge Doubling Curve*, a graphic representation of this growth. It shows that it took 1500 years for the sum of human knowledge first to double in size from where it was around 1 CE; after that it doubled again approximately every 100 years until the twentieth century [2]. Today this doubling is reckoned to take place every 12 months, and some almost apocalyptic estimates claim that it is likely to happen every 12 hours in the near future [3]. Hierarchy seems doomed in the face of this onslaught. But what other ways are there?

Flattening the Hierarchies: Facetted Classification

One approach that comes from the world of libraries is not to abandon all hierarchy but to flatten it to an absolute minimum. Here, once again, we meet the great Indian librarian, S.R. Ranganathan. He initially trained as a mathematician, only changing course to turn his logical mind to the fundamental theories of information science in

R. Gartner, *Metadata*, DOI 10.1007/978-3-319-40893-4_7

his 30s when he started his two decades-long tenure as Librarian at the University of Madras (now Chennai). The inflexibility of enumerative classification was one of the first issues to which he turned his critical eye.

The major problem for him was the closed picture of knowledge that this approach required:-

> An enumerative scheme with a superficial foundation can be suitable and even economical for a closed system of knowledge……What distinguishes the universe of current knowledge is that it is a dynamical continuum. It is ever growing; new branches may stem from any of its infinity of points at any time; they are unknowable at present. They can not therefore be enumerated here and now; nor can they be anticipated, their filiations can be determined only after they appear. [4]

Enumerative systems, he claims, are inherently superficial and out-of-date from the moment they appear. His response to these challenges was his renowned (but not often implemented) Colon Classification, which we met briefly in Chap. 2. In it he introduced a new approach to organizing knowledge known as a *facetted* classification.

Back to etymology once more: the French word *facette* from which the English *facet* derives is the diminutive term for face. This 'small face' referred initially to the cut face of a diamond [5]. Ranganathan is usually credited as the first to apply it to the field of information science. In his grandly named *Prologomena to Library Classification*, he refers to it as "a generic term used to denote any component – be it a basic subject or an isolate – of a Compound Subject" [6], in other words an atomic concept from which more complex ones can be created by combination. Crucially, there is no hierarchy implied between facets when a cataloguer uses them to define a compound subject: they are considered equal from this perspective.

The use of facets may flatten the hierarchies of an enumerative system but they do not necessarily eliminate them entirely. Most facetted schemes group them together into 'classes': Ranganathan himself proposed 31 of these, all broad categories (such as history, medicine and literature) of the type that appear at the top of Dewey's scheme. Unlike Dewey's dense tree-like structure, this hierarchy stops at two levels only: the classes and the facets which are sub-classes of these. A crucial feature is that each facet should be exclusive, not overlapping with its neighbours. Clarity, clear semantic boundaries and full coverage of all aspects of its parent class are the aims when constructing a facetted classification.

As in all areas of metadata, these grand ideals are harder to realize in practice. The problem when implementing the Colon Classification in a physical library can be the tortuous shelfmarks that the scheme produces. This often-quoted example comes from Ranganathan itself:-

L,45;421:6;253:f.44'N5

This lengthy string translates to "research in the cure of tuberculosis of lungs by X-ray conducted in India in 1950" (Fig. 7.1).

The boxes at the top, to the right of the main class, represent what Ranganathan terms his five 'primary categories'; these are used to order the facets that make up a

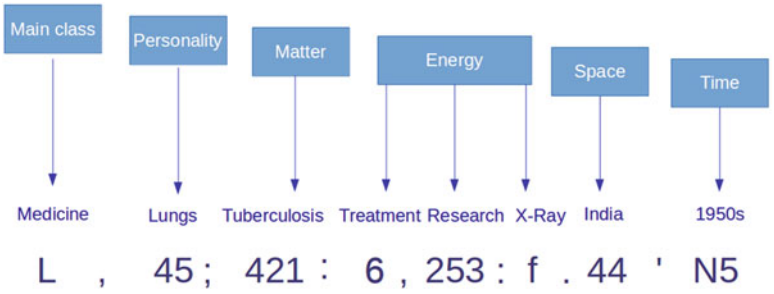


Fig. 7.1 A Colon Classification shelfmark parsed into its component facets

subject. These categories, Personality, Matter, Energy, Space and Time, are usually known by the acronym PMEST. They are semantically wide but clear to identify. Personality is a subject's distinguishing characteristic, what it is that allows us to tell it apart from another to which it is closely-related. Matter is any physical dimension to a subject, the physical material of which it is composed, Energy any action that takes place in relation to it, Space its geographic location and Time its temporal span.

The rationale for having these categories is the pragmatic one that in the real world, dealing with physical objects such as a book or journal article, there has to be some order to the arrangement of facets if something as prosaic as a shelfmark is to be created. The sequencing of the components of a compound subject is known technically as a *citation order:* knowing something of the logic behind it in a facetted scheme is essential for the cataloguer and also helpful to the user if they are going to find their way around the subjects labelled with these lengthy strings.

Unfortunately, this is something that has often proved too much to ask of the general user, who is likely to find the PMEST citation order rather abstract and esoteric. Certainly, it almost always proves harder for them than an enumerative classification such as Dewey's, the logic of which is readily apparent even to children when they first encounter it in their public library. Expressive and flexible facetted classification may be, but it has not found its way into libraries to anything near the extent of Dewey's venerable scheme.

In the online world, things are rather different. Here the iron rule of the citation order is not required as it is easy for us to choose and manipulate facets when we carry out a search. Take the example of the British Library's online catalogue. When I put in some keywords to find a book, the matches it finds are accompanied by a number of options to refine my search, to cut it down using several types of facet (Fig. 7.2).

The expandable menus on the left show the categories within which these facets are grouped, including their material type (books, articles, audio), subject, author, language and so on. Clicking on one of these reduces the list of matches to those that share the facet chosen. In this way, a long list of matches, often numbering tens of thousands for broad search terms, can be whittled down very quickly and effectively to manageable proportions.

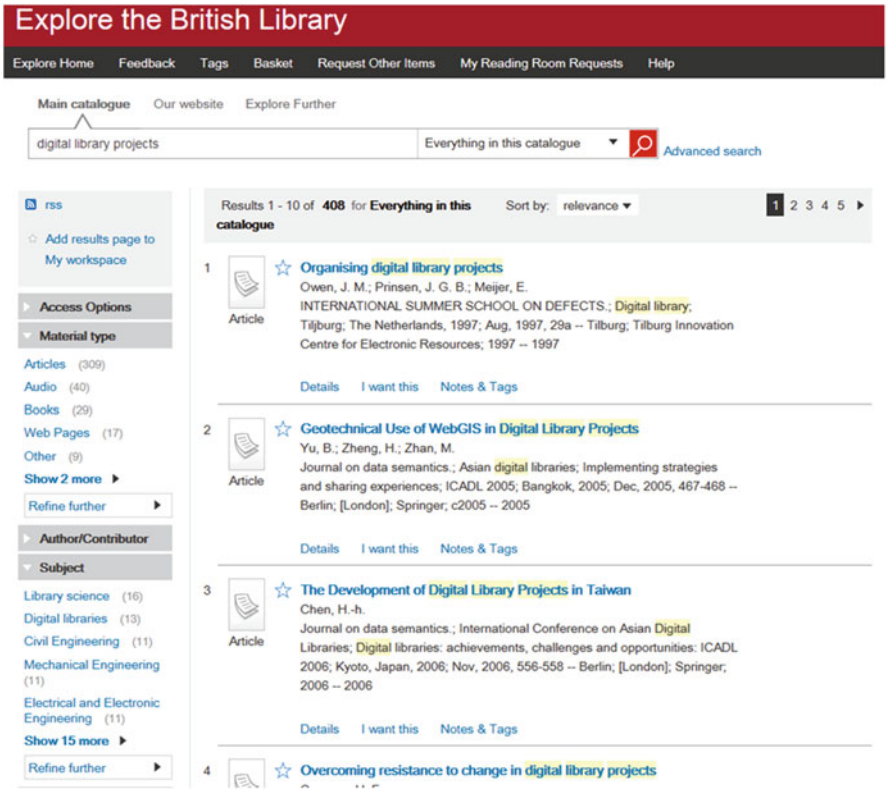


Fig. 7.2 Facetted browsing in the British Library Catalogue © The British Library

This method of facetted browsing is very popular in e-commerce sites as a way of allowing potential shoppers to refine their choices before parting with their money. Amazon, eBay and many others use it to guide buyers to what they want. The electronic medium has allowed the potential of Ranganathan's approach of facetted classification to operate without the straitjacket of citation order needed in its physical counterpart. In a sense, and perhaps ironically, it is with Amazon and its competitors that his vision has at last been realized.

Great claims have been made for facetted classification: it has almost attained something of a cult status in the world of library science. Certainly it has introduced a degree of flexibility into classification which enumerative schemes struggle to achieve although at the cost of greater complexity when it comes to deciphering the abstruse classification codes it generates. It has also lost something of the simple interoperative power of schemes such as Dewey's. A Dewey number is a Dewey number is a Dewey number wherever it is found and one subject should have a single, unambiguous number attached to it. In a facetted scheme, the same subject in one library may be labelled with a different code in another, depending on the judgment exercised by each cataloguer when combining the facets available to them.

Of course, compromise is possible here and schemes can combine enumerative and facetted approaches, the first to lay out the overall ground plan for a classification, the second to allow flexibility and growth. Even the venerable Dewey scheme, the *doyenne* of enumerative classification, incorporates facetted elements to allow this. Any Dewey number may, for instance, be suffixed by a geographic code to show the part of the world to which it applies. Other facets may cover such add-ons as languages, historical periods and ethnic or national groups. This is a pragmatic solution which works well in squaring the inflexible circle of enumeration into something more usable.

Ontologies: Metadata as a Network

Facetted classification goes only so far in removing subjects from the constraints of the hierarchical model. It flattens the hierarchy to fewer levels and allows much more flexibility than the implacable logic of an enumerative scheme, but it usually involves some layering, either in the relations between facets and classes or in the citation order within which they are arranged. A more fundamental reordering of the organization of knowledge has occurred in the world of information science since the early 1990s. This new way has assumed the rather grandiose name *ontology.*

In philosophy 'ontology' refers to the study of the nature of being or existence. Its own existence goes back over two and a half millennia to ancient Greece, where such great minds as Plato, Aristotle and Parmenides all brought their thoughts to bear on such questions as "what is existence?", "when can something be said to exist?" and "what is a thing?". Much later, St. Thomas Aquinas introduced the notion of the *ens qua ens,* ("being as being") which later philosophers such as Christian Wolff and Gottfried Leibniz adopted as the primary focus of their philosophical investigations. Immanuel Kant turned his mind to it in the eighteenth century, forcing us to acknowledge the limits of our capacity to understand ontology because of the limits of our cognition. And in the twentieth century, Martin Heidegger pondered the subtle ontological difference between being and Being in his magnum opus *Being and Time* [7] .

With so many centuries of thought lying behind it, the notion of ontology with which we are concerned here may certainly seem prosaic. Prosaic or not, it was first propounded by the computer scientist Tom Gruber (who later achieved fame as the creator of Siri, the talking interface to the iPhone) in an article of 1990. He defined an ontology as a "specification of a contextualization", and a contextualization as "an abstract, simplified view of the world that we wish to represent for some purpose" [8]. So far it is hard to see how this differs from what a standard taxonomy is trying to do. The crucial difference is the way in which an ontology brings its components together. Gone are the rigid rungs of the enumerative or even facetted ladder: now concepts can be related to each other like the strands of a spider's web, a network of ideas and linkages that can be as flexible as the subject being modelled requires.

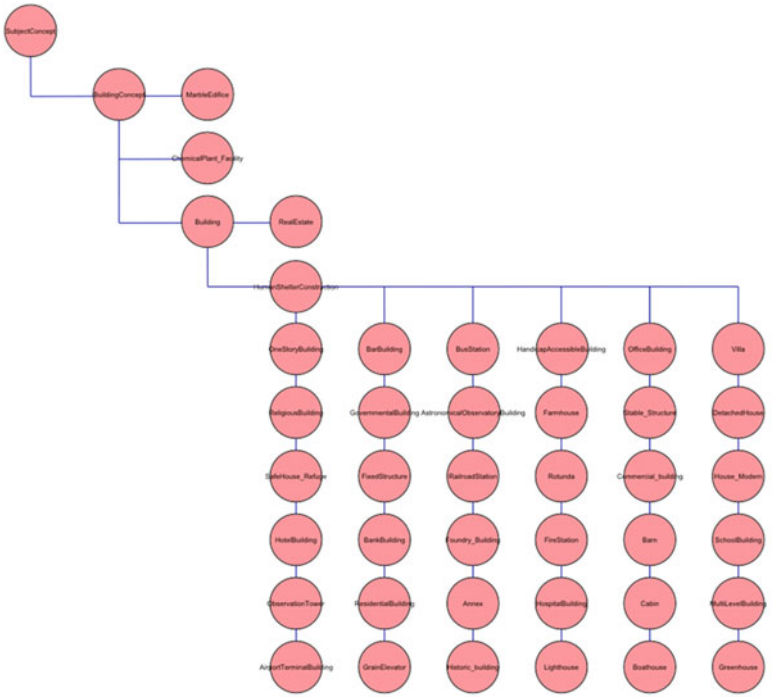


Fig. 7.3 The architecture of a hierarchical taxonomy (Diagram by Michael Bergman)

What we have here is a move from this arrangement (Fig. 7.3) to something like this (Fig. 7.4).

One feature of this arrangement that is immediately obvious is that each component or concept can be connected to any number of its companions: gone is the requirement for it to have a single slot in the overall enumerative tree. This means that it becomes possible to navigate between concepts much more flexibly. No more is it necessary to move up and down the hierarchy (or possibly sideways at the same level) to move around this web of concepts, the routes taken can now be as circuitous as the network of connections that the ontology defines. Nor is it necessary to begin a journey through the concepts at any pre-designated starting-point (such as the top level in a hierarchy): you can enter the web at any of these points and follow any route that leads from there.

This sounds liberating compared to the strictures of a scheme such as Dewey's but it is by no means a metadata free-for-all. You do have to follow the connections prescribed by the designer of the ontology and these may be as flexible or rigid as they see fit. It is perfectly possible to design a scheme as unyielding as Dewey's within an ontology structure and to use it in the same way as its more traditional counterparts. But most designers prefer to embrace the new possibilities of this approach and design more accommodating networks of information that are less prescriptive.

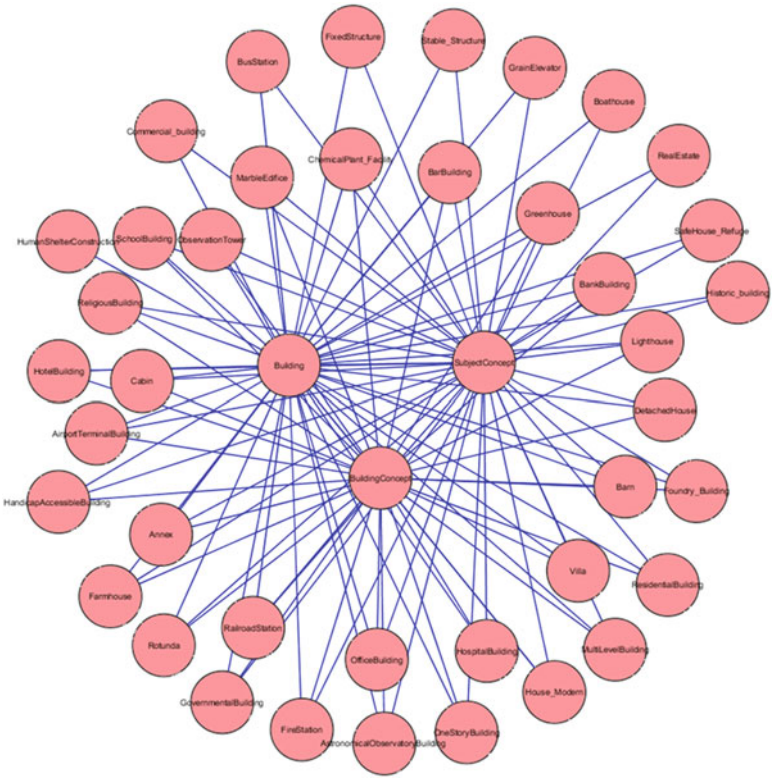


Fig. 7.4 The networked architecture of an ontology (Diagram by Michael Bergman)

The key rationale behind taking this approach is that it can be considered a more accurate way of representing the structures of knowledge as we humans perceive it. Hierarchies may appear to be innate to human thinking in many ways but we know from our own experience that the workings of the human mind are much messier than these tidy structures may suggest. The way we construct knowledge inevitably reflects something of this mess, despite the tidy models we saw earlier that view it as a pyramid built upwards from a ground level of data, each stage representing a type of inverse entropy that is increasingly ordered compared to the one below. Perhaps the ontology model allows us to preserve the messy human experience of knowledge while allowing us to shape it through metadata and so mould it from our data and information.

This is an assertion often made by ontologists who see it as a great step forward in metadata. A significant advantage often claimed for ontologies is that they can be used to draw inferences that are not explicitly present in them when they are compiled. This can be done by adopting the simple techniques of logic which form the part of any philosophy course and using computers to apply them. To do this requires that the links within an ontology make statements that themselves carry meaning.

The links in an enumerative scheme are not, of course, meaningless. They tell us that a concept within the hierarchy is a type, part or instance of the concept immediately above it. Similarly, its links to concepts below tells us that these are types, parts or instances of the concept itself. Often these definitions are left at this semantically imprecise level. As human beings, we can intuit roughly what we mean by 'is a type of' or 'is a part of' and the hierarchy as a whole helps up work this out. When we do not have the support of a hierarchy of concepts, we need to be much clearer about what the linkages mean. Part of defining an ontology is defining the semantics of these links.

So what do we find when we look into an ontology? First and foremost are the concepts themselves. These are usually called *classes* in ontology-speak. Often they are divided into *sub-classes*, which as the name implies are more specific concepts than their parent superclass. Within these classes and subclasses can be located *instances,* the individual objects, concepts or things which have the properties described by their parent. To make these ideas more concrete, or perhaps more dough-like, we could look at how they might appear in a simple ontology to describe the features of a pizza (one that many an ontology-rookie constructs to learn the ropes). Here the top-level class may be called *pizza*, and be subdivided into sub-classes, *meat pizza* and *vegetarian pizza*. Within the *meat pizza* sub-class we might find such instances as *pepperoni pizza* or *spicy beef pizza* and within the *vegetarian pizza* sub-class such favourites as *pizza margherita* and *four cheeses*.

Immediately alarm bells should go off here: we are back in the realm of hierarchies, in the form of classes, sub-classes and instances. This is one of the ironies of the ontology model: most impose a tree-like hierarchy for their internal structuring as rigid as that found in many an enumerative scheme. The crucial difference is that this design feature is optional (although given its pervasive presence in ontologies it is often difficult to remember this) and can be ignored if desired. It would be quite possible to omit the higher level classification in this pizza ontology and stick to the pizza instances themselves, although the convenience of classes and sub-classes usually overrides such an urge. What does single out an ontology from an enumerative approach is an array of features that allow us to describe precisely the semantic features of a concept in itself and its relations to its peers.

This is done by assigning *properties* to classes, sub-classes or instances; these are semantic statements which can be understood by a computer. We might want to say that the *pizza* class has two properties, one that it must have a base (we could call this *hasBase*) and one that it must have a topping (*hasTopping*, perhaps). A property such as *hasTopping* can then be linked to a class of pizza toppings in which the properties of each are defined. We can then link from instances of a pizza (such as a margherita) via the *hasTopping* property to an instance of a topping (*cheese* or *tomato*) and then to the properties associated with the topping itself (such as whether it contains meat or not). Ontologies also allow us to specify whether we should have an exact number of a particular topping (the Four Cheese pizza will require this to be set to four), and whether all of them should be of a specified type (*cheese* in this case). Very quickly a complex web of semantic links can be built up.

One further feature which allows a more sophisticated engagement with metadata than a conventional classification is that ontologies allow us to apply restrictions on properties. These can be in the form of exclusions, stating that if one applies, another cannot be valid: if our topping has the property *containsMeat,* it cannot also have the property *isVegetarian*. These can be powerful mechanisms when trying to analyze large and complex ontologies.

All of these features allow sophisticated inferences to be drawn by applying simple logical tests to this web of classes, sub-classes, instances and properties. I may not have explicitly labelled a margherita pizza as vegetarian in my ontology (perhaps by including it in a sub-class called *vegetarian pizza*), but if I have set up its linkages correctly (specifically not including any topping with the property *containsMeat* in its ingredients) it should be easy to infer that this is the case. Sophisticated software packages, known as inference engines, can interrogate complex ontologies to draw conclusions of this type.

The construction of ontologies may appear an esoteric discipline and perhaps a lot of effort to determine something as uncomplicated as whether a pizza is vegetarian or not. There are, however, some ontologies that are gaining traction for more practical purposes. One well-known example that has become relatively popular is called *Friend of a Friend (FOAF)* [9], a simple ontology for recording basic information on people and their connections to others. The details that may be stored here include their interests, education, and workplace; they can also incorporate links to other people whom they know in any way.

This ontology can be thought of as one of the first attempts to construct the type of social networks we now take for granted in such giants as Facebook (which it predates by 5 years). It has not achieved anything like the ubiquity of that social media platform, however. A few web browsers, including Google Chrome and Firefox, have made plugins available which can detect and point to FOAF metadata and social blogging sites such as LiveJournal allow their authors to include profiles encoded in FOAF. But in many ways the abstruse method of using an ontology to define a social network has been overtaken by more user-friendly social media sites which hide their complex background metadata from the user.

More traditional metadata has also found its way into this world of networked knowledge. A generic ontology for constructing thesauri and classification schemes within these more flexible structures has appeared in recent years. Called, perhaps a little optimistically, the *Simple Knowledge Organisation System (SKOS)* [10], it is a framework for constructing controlled vocabularies of any type. As with any classification scheme, a SKOS vocabulary has at its centre the core concepts that represent the ideas, objects or subjects that it is constructed to express. To these can be attached properties that define their relationships to others of their kind, including such familiar acquaintances as *broader*, *narrower* or *related;* more powerful is the property *semantic relation* which indicates a relationship that means something more specific than these borrowings from the structure of a conventional thesaurus.

This ontology also allows us to define the degree of semantic proximity between terms, to say whether one concept is an exact match for another, a close one or a

distant one. This can be very useful for enabling 'fuzzy' matching when searching SKOS metadata; it makes it possible for us to find information that would otherwise be missed using the more precise criterion of exact matches alone. In some ways, this allows us to make our information more human, closer to knowledge as we often encounter it. Rarely is our interaction with knowledge as clean-cut as a search algorithm would like it to be and acknowledging this in our ontologies can be regarded as a more honest approach to modelling it.

SKOS is definitely an ontology with great potential, in some ways the obvious way to allow the established practices of metadata organization to evolve into the technological environment of the early twenty-first century. Its adoption remains, however, embryonic at the time of writing. Some important controlled vocabularies have been published in SKOS, including the *Library of Congress Subject Headings*. A large geographic database, *Geonames* [11], a listing of over ten million place names which forms the basis of many web services, also uses SKOS for its underlying semantics.

These are all significant applications, although they remain at present small in number. Part of the problem is undoubtedly that the world of ontologies can seem intimidating and much wrapped-up in geek-speak. For those who work in areas such as libraries, which have long-established practices well within the comfort zone of their practitioners, ontologies often appear to sit more readily in the realm of the computer or information scientist. They may well appear too complex for what is needed to run a library or compile a bibliography. The truth is that they can be as simple or complex as is required and what they are attempting to do is what those who construct metadata have always been seeking to achieve.

References

1. Library of Quotes. (2012). Hierarchy works well in a stable environment. *Mary Douglas*. http://www.libraryofquotes.com/quote/837561. Accessed 29 Oct 2015.
2. Fuller, R. B. (1981). *Critical path*. New York: St. Martin's Press.
3. Industry Tap. (2013). Knowledge doubling every 12 months, soon to be every 12 hours. *Industry Tap*. http://www.industrytap.com/knowledge-doubling-every-12-months-soon-to-be-every-12-hours/3950. Accessed 1 Feb 2016.
4. Ranganathan, S. R. (1951). *Philosophy of library classification*. New York: Hafner.
5. Harper, D. (2015). *Online etymology dictionary*. http://www.etymonline.com/index.php?term=facet. Accessed 7 Oct 2015.
6. Ranganathan, S. R. (1957). *Prolegomena to library classification*. London: Library Association.
7. Heidegger, M. (1962). *Being and time*. New York: Harper.
8. Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing? *International Journal of Human Computer Studies, 43*, 907–928.
9. Brickley, D., & Miller, L. (2015). The FOAF Project. http://www.foaf-project.org/. Accessed 29 Oct 2015.
10. W3C Consortium. (2012). SKOS Simple Knowledge Organization System – home page. http://www.w3.org/2004/02/skos/. Accessed 29 Oct 2015.
11. GeoNames. (2015). GeoNames. http://www.geonames.org/. Accessed 1 Feb 2016.

Chapter 8
Breaking the Silos

Underneath the Jura mountains, on the border between France and Switzerland, lie the tunnels that house the particle accelerators of CERN, the European Organization for Nuclear Research. Within these subterranean lairs, high energy particle beams are accelerated to near the speed of light and collided to bring into a transitory existence such elusive creatures as the Higgs boson, the elementary particle whose associated field gives mass to most of its counterparts. It is cutting-edge physics that take place beneath the landscape of the Jura but it impinges less on the imagination of most of the world than an innovation that had its humble origins in the offices of one of CERN's fellows, Tim Berners-Lee.

It was 1989 that Berners-Lee came up with a novel solution to the problem of managing the huge amount of documentation that the operations at CERN were generating every day. He particularly saw the potential of a technique known as *hypertext,* a way of moving between documents by embedding links within the text they contained. Any word or phrase in a hypertext file can be made a 'hot link': by choosing it, the user is taken straight to another file that the link points to. It was a quick and easy way to get a grip on a confusing mass of information and make it more manageable as it continued to grow.

Berners-Lee put together a simple graphical interface to his hypertext system which he had by then given the name World Wide Web. This operated much like the more sophisticated graphical browsers of today, such as Firefox or Chrome, allowing users to click on a hyperlink and load the document that perched at its other end. Technical reasons prevented others outside CERN from using the interface and so the epithet 'World Wide' might have seemed a little presumptuous at this stage. But it soon managed to spread beyond the confines of its parent institution in the form of a basic line-mode browser designed by a student named Nicola Pellow (Fig. 8.1) [1].

This appears an unassuming start to something that has changed the world as profoundly as the Web. Simple as it may seem now, this browser was doing much the same as its present-day successors that we know so well, joining together documents or other objects by embedded hyperlinks. Berners-Lee's most inspired idea

R. Gartner, *Metadata*, DOI 10.1007/978-3-319-40893-4_8

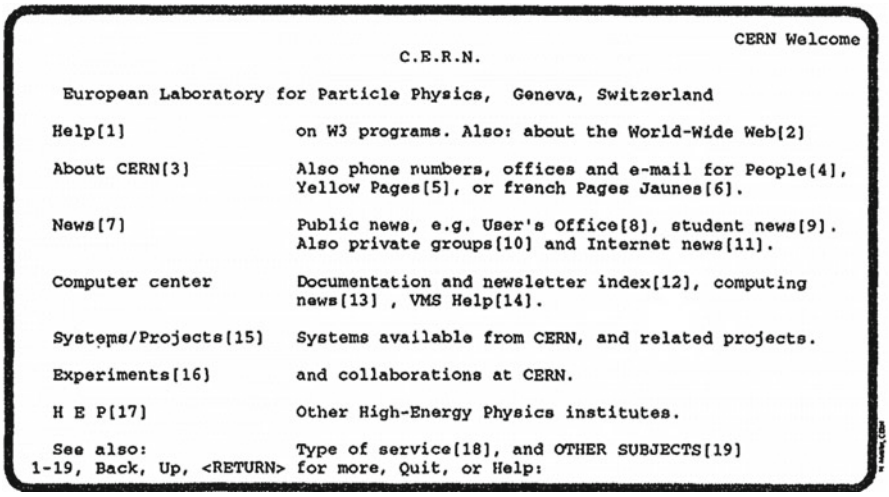


Fig. 8.1 The simple browser that introduced the World Wide Web to the world (image © CERN)

was to use Uniform Resource Identifiers (URIs) to record the address of the objects on the Web so that they could be referenced in this way.

CERN already had in place a method for tagging its documentation using a markup language known as SGML (Standard Generalised Markup Language). SGML is the precursor to XML, rather more complex (and so powerful) but harder to use and process: it was to simplify its use that XML was to be devised several years later in 1996. To allow hyperlinking, Berners-Lee added a new element, <a> (anchor), to the small set of tags already in use within CERN. He in effect created a new SGML application which he termed HTML (Hypertext Markup Language), actually not a new language at all but an application of an existing one. Despite this misnomer, it was from such a small digital acorn that the digital oak of today's World Wide Web has grown.

From Documents to 'Things'

Berners-Lee's innovation justly made him famous: he had effectively broken the silos housing the world's data by allowing them to link up in this way. But this was far from the end of his ambitions as he had something much grander in mind for his Web. What did its linkages tell us? Merely that at the end of each was a document or some other digital object. Pretty banal stuff. Much more interesting would be the prospect of the links telling up something of what they themselves mean. As it stood, a link on the Web that pointed to something said nothing more than "here it is"; what if it could say "this is what it means"?

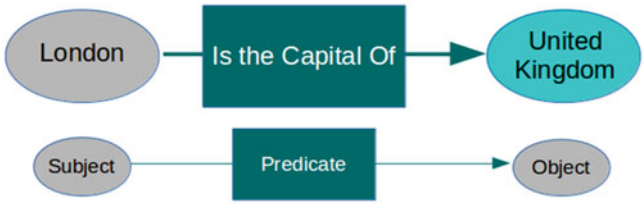


Fig. 8.2 An RDF 'triple'

He and two colleagues published their ideas in an article in *Scientific American* in 2001 in which they christened their reborn vision the *Semantic Web* [2]. This Web would become 'semantic' in the sense the links that made it up would have meaning embedded within them and this meaning would be something that computers could make sense of and manipulate in intelligent ways. This would not be a new Web, they claimed, but rather "an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation" [2].

To do this required new ways of encoding information and the metadata that would provide the links between its components. Some way of recording the semantics of these linkages was necessary that could turn them into the type of metadata that computers could process. To do this, Berners-Lee and his colleagues suggested using a relatively new method for expressing semantic links known as the *Resource Description Framework* or *RDF*.

RDF is a way of expressing the meaning of a connection between two 'things', abstract or physical. It is structured like a very simple sentence containing a subject, a predicate (a verb and any words that the verb governs) and an object. An RDF statement could use these three ingredients to make an assertion about geography (Fig. 8.2).

Almost any metadata (and data) can be split up into these small molecules of information. Because of their tri-partite structure, they are usually referred to as 'RDF-triples' or often simply as 'triples'.

In this example, the triple's three components are shown as strings of text: these are easy for a human to read but somewhat vague for the more prosaic thought processes of a computer. Which 'London' are we referring to here – London (Greater London), London (Ontario) or one of the eight cities with that name in the United States? What exactly is meant by 'Is the Capital Of' – a capital city, an uppercase letter of the alphabet or someone's accumulated wealth? In most cases, our linguistic and reasoning skills allow us to work out the sense of a statement such as this, but computers need more help in telling these fuzzy areas apart.

Help is at hand in the form of the URI, that ubiquitous identifier for everything on the Web. The same triple could be encoded in something like the form in Fig. 8.3.

Here, the strings 'London' and 'United Kingdom' are replaced with URIs from the *Getty Thesaurus of Geographic Names*, a vocabulary of place names, and the predicate 'Is The Capital Of' by one from a small ontology of geographic terms. This translation into URIs makes it harder for humans to read and understand but

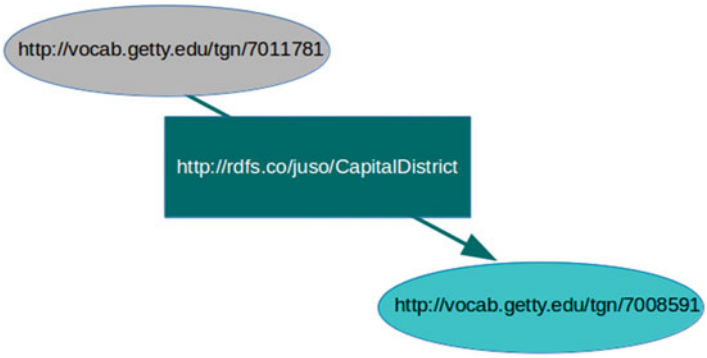


Fig. 8.3 The same triple encoded using URIs

much easier for machines to process. Using these URIs means that there is no need to rely on human reasoning, intuition, memory or cognitive skills to know exactly what is being asserted in this triple.

Berners-Lee's vision was that RDF triples would form the atoms (or more appositely given their compound nature) the molecules on which the Semantic Web would be constructed. Billions of them would enable the entire Web to become one universal repository of structured data that could be searched, manipulated and processed as if it were one database. The Semantic Web would have no centre but, like the World Wide Web itself, would take its shape from its contents and change continuously as they do.

There are two ways in which the Semantic Web 'breaks the silos' in the world of metadata. First of all, it blurs the boundaries between metadata and data, always slightly fuzzy but in the context of networks of triples more difficult to draw as a solid line. Metadata that is neatly wrapped up in a standard is easy to distinguish from the data it describes if that data is itself packaged into a discrete object with clear boundaries separating it from its neighbours. We can readily tell a MARC record from the book it refers to because of this packaging. When both form part of the same merged database it can be harder to distinguish them. We will speak of data throughout this chapter but for these reasons this should always be taken as including metadata as well.

Secondly, it can smudge the edges of sets of data to another hazy blur. Although we can still identify the separate datasets that make it up, the whole Web can now supposedly be treated as a single database and function as one. If one part of it is isolated from the others, it is rendered pointless. We can now easily cross the borders between collections of data as if they did not exist: to maintain barriers between them would be to nullify the *raison-d'etre* of the whole enterprise.

One hurdle to realizing this vision is the problem of consistency in this morass of data produced without any centralized codes of practice. Berners-Lee points out in his original article that different URIs can readily be used for the same concept (for instance, for the city of London): which one should we use? This is where ontologies, the subject of the previous chapter, come into their own. They can provide

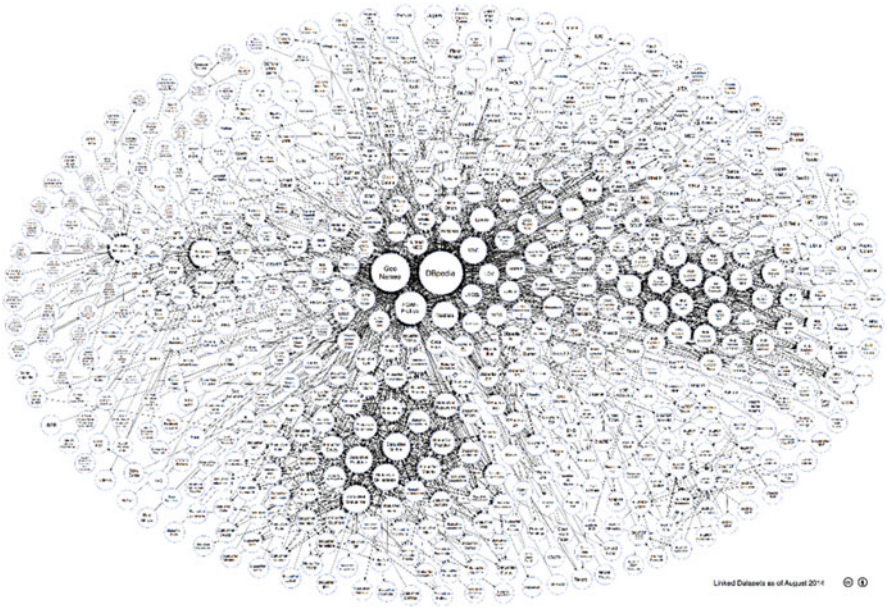


Fig. 8.4 Linking Open Data Cloud Diagram 2014, by Max Schmachtenberg, Christian Bizer, Anja Jentzsch and Richard Cyganiak. http://lod-cloud.net/

ways of reconciling inconsistencies of this type, mapping URIs to each other to show that they refer to the same thing and so effectively tidying up the mess of uncoordinated data. They can do far more than this, of course. They can provide structure to the Semantic Web through the semantic relationships they encode and so allow machines to make inferences about what is out there on scales much larger than a single set of data would allow.

Berners-Lee's vision is a grand one and one that makes sense technically. It has proved inspirational enough for many to embrace its principles. Opening up data is now a high priority for many and the primary mechanism to do this is to expose it to the world as triples – when this is done, it is usually called *Linked Open Data (LOD)* A iconic image, known as the *Linking Open Data Cloud Diagram,* is often used to demonstrate how far this phenomenon has reached. Figure 8.4 shows the state of this 'cloud' in April 2014.

It is a dense, thickly-populated mass that is squeezed into this oval: each circle within it is a collection of data that is the fruit of the work of many people. Here we find data from such august bodies as the BBC, the Library of Congress, government bodies, research agencies, the list goes on. At the centre is *DBPedia*: this is an initiative by the founders of *Wikipedia* who publish information extracted from their online encyclopedia as LOD under this name [3]. It occupies this prime position in the cloud because it is so frequently the first port of call for links from other datasets. If the world of LOD has anything approaching a centre, this multi-domain dataset is probably the best contender for this distinction.

The links between these datasets form a network so dense that it is almost impossible to decipher. This certainly looks like a web that is vibrant and growing, but it cannot as yet claim to be a genuinely global Semantic Web: impressive as it looks only a small part of the world's data is there. There is definitely a whiff of disappointment amongst advocates for the Semantic Web: why hasn't it grown as exponentially as its less meaningful forebear?

As far back as 2006 Berners-Lee expressed something of this disappointment. In a co-written article entitled *The Semantic Web Revisited*, he admitted that despite some notable successes, including initiatives in government, defence, business and commerce, it was "still apparent that the Semantic Web isn't yet with us on any scale" [4]. At the time of this article, many of the technical features needed to make it work, including the development of triple stores capable of handling the billions of triples needed, were still taking shape. Almost 10 years on, with these problems effectively solved, the honest opinion of many of those proselytizing on behalf of the Semantic Web would be similarly downbeat: it hasn't gone viral, it hasn't accumulated enough critical mass to make it indispensable.

There are many reasons for this, some of which Berners-Lee points out in his article. Managing data encoded in RDF triples can be much more complex than when it is neatly packaged in discrete bundles. Information that can be encoded neatly and concisely in a structured format is spread much more thinly and is harder to disentangle when converted to a set of triples. In 'triplifying' information we are increasing its entropy, its state of disorder. The same information is there and can be reconstructed from its fragmented state, but it is harder to deal with. We can take a patterned plate, smash it, and still work out the pattern from the remnants, but it is much easier to work with the plate in its original low-entropy, intact state. So it is with data.

The blurring of boundaries, one of the great strengths of the Semantic Web philosophy, is also one of its key drawbacks for those who have to look after the data that has been 'triplified'. The difficulty is most acute for those looking into the future who want their data to be usable to their descendants. Digital preservation, the long-term curation of data, is well established as a discipline and has codified sound principles which can make us reasonably confident that the data we want to preserve can be preserved. Most of its practices depend on grouping data and metadata into neat packages with clear boundaries: it is these packages that we archive in strictly controlled ways when we want to ensure their longevity. When these boundaries are indistinct and when we don't know where the edges of our packages lie this become much more difficult. When the whole Semantic Web is potentially one giant database, where does the responsibility for preserving our own data stop and someone else's take over?

The same problems apply when it comes to defining intellectual property and protecting the rights of those whose hard work has produced the data. How can we define what is ours and what is someone else's work? The copyright of a discrete package, such as this book, is easy to assert when we know its boundaries: but if its contents are not a neatly-bound volume or a distinct digital object such as a PDF file but instead millions of triples, which only 'work' if they interact with others in the

Semantic Web, how can its copyright be defined and protected? These are not insuperable problems, but they present challenges in a legal area that is so heavily predicated on boundaries to define the objects it covers.

For those who want to use the data as opposed to manage it, one of the most pressing issues that Berners-Lee admitted to is provenance and trust [4]; the first of these is knowing where the data we find comes from, the second is using this knowledge to work out whether we can have confidence in it. A neatly packaged blob of data is generally quite easy to trace if its owners want us to do so: but the origins of an amorphous collection of triples from diverse sources whose boundaries are not clear are much more opaque. A set of triples from the same source can, of course, be labelled with their provenance and we can then work out whether we can trust them based on this evidence. But when they intermingle with others and our knowledge is formed from the sum of these (or if inference is working, more than the sum) it becomes much harder to establish confidence in them.

All of this sounds very negative but none of these problems are insuperable. Most can be resolved by combining the technical possibilities of LOD with changes to cultural attitudes to data and to the mechanisms long established to handle the comfortable containers of old which are now beginning to show their age. Above all, this data needs applications that use it in ways that do not require any knowledge of RDF and its arcane workings. There have been some brave efforts in this area. The BBC is one body that has had an active LOD programme since 2010; it has used it for its news and sports websites which benefit from its flexibility and ability to bring together data in real time from multiple sources [5]. But most who make their data available in this form stop short of providing approachable interfaces to it. Their data may be open but its accessibility to the lay user can be another matter.

How insuperable these problems will prove is hard to guess. Interest in the Semantic Web and its possibilities seems to have ebbed and flowed since it was first mooted, though many of its proponents retain a proselytizing enthusiasm for it. It certainly has the potential to break down the silos between data and metadata and between data and data. It has the great advantage of allowing us to do this in a structured way so that we can still find our way through the soup of triples that it serves up. Despite there being a distinct sense of running out of steam, it is hard not to feel that the Semantic Web is just awaiting its next big idea to make the impact it has for so long dangled tantalizingly before us.

The Metadata Bypass: Content-Based Retrieval

Berners-Lee's Semantic Web offers an ingenious way to give meaning to the Web and make it more than the sum of its scattered parts. His vision is of silos of data fading away under the inexorable force of URIs and semantic links. But his approach is not the only one we could conceive of achieving this. Another was already well-established by the time he wrote his seminal article, one which bypasses any hint of metadata created by human hand and goes straight to the contents of websites

themselves. This way of letting us get at what the Web has to offer is known by the inelegant but entirely apt term *content-based retrieval*.

Today we know this type of retrieval in the form of the search-engine giant Google and its smaller rival Bing. For its first decade of operation, Google was in the company of a small circle of rivals, including AlltheWeb and AltaVista, all now sadly defunct. These certainly seemed, and still seem in case of the survivors, to have realized something of the vision of a single Web that can be searched as one set of data, all without the clever but esoteric semantic glue of RDF and ontologies. They certainly appear to have achieved the aim of breaking the data silos without the assistance of metadata or any other human intervention. Could this make thousands of years of endeavour in defining and creating 'data about data' irrelevant?

Content-based retrieval is trying to do much the same as descriptive metadata, help us find what is out there and work out if it is useful or relevant to us. Both approaches try to abstract something of what an object or document is 'about' and match it against another 'about': this second 'about' is the subject of our search, whatever it is we are looking for. The more similar the two 'abouts', the closer the match and the more relevant should be the object they describe. The key difference between content- and metadata-based retrieval is that the former uses mathematics to build up the abstracted 'about,' forsaking the human intermediary of the latter who creates metadata to achieve a similar end.

The actual algorithms used by Google or Bing are closely-guarded secrets and are constantly being refined but the overall principles that content-based retrieval follows are simple enough. These systems create indexes, akin to those found in the back of a book, which put the words contained in documents into a convenient order that allows them to be retrieved efficiently. The words in the index point us to the documents in which they appear in the same way that an entry in a book index points to a page number.

This has only limited use when we are dealing with the whole Internet: finding out that 200 million documents contain a word such as 'computer' is not going to be of much use to anyone. We have to sift through these by putting them into some order of importance or relevance. This can be done by assigning weights to each word depending on how important it is likely to be to our search. Word frequency, how often a word appears in a document or a body of documents (often called a *corpus*) is the key to assigning these weights and so letting our searches pull out the most relevant results.

Generally, the assumption is that the more often a word appears in a document the more important it is within it. This makes sense intuitively, but we soon run into trouble if we follow this approach blindly. Words such as 'a'. 'the', 'it', 'of' and so on will pepper every document but tell us nothing about its content. We can get round this problem by creating dictionaries of these useless words, known as stop lists, and excluding them from our searches. But this will still leave us with plenty of words which turn up often but tell us little. Just how irrelevant they are may not be immediately obvious: this will vary according to their context and working this out without human intervention can be difficult.

Plenty of studies have been conducted around these questions, and the majority seem to show that the most useful words in a document tend to be those that lie in the middle frequencies: the most common are usually irrelevant bits and pieces that should go into stop lists and those that rarely appear are generally insignificant. So most content-based retrieval concentrates on this middle ground. This alone only goes so far when it comes to ranking millions of documents. What is needed on top of this is some sense of how specific the terms are: more precise ones, those with a narrower semantic scope, are likely to be more important than those which are broader. If we use a very narrow term in our search and we find it in a document, we can be pretty sure that we have found what we are looking for.

But how to work out how specific a word is without human intervention? This is where we look at the corpus as a whole, the entire collection of documents that we are searching (potentially the whole Internet). The assumption here is that the *less* often a word appears in the corpus *as a whole*, the more specific it is likely to be. If it turns up often in a single document but doesn't appear anywhere else in any numbers, we can be sure that it is term of high precision and so should give it plenty of weight when we rank the relevance of the document we have found.

So content-based retrieval balances two opposing principles: a term gets more weight if it appears frequently in a single document but less if it appears commonly in the whole corpus. Usually its weight is calculated by multiplying its frequency in a document by the *inverse* of its frequency in the corpus as a whole (which gets larger as its frequency gets smaller). This weighting can then be used to rank the documents we find.

This is an oversimplification of enormous magnitude but it illustrates how statistical techniques can begin to replicate the processes human beings bring to deciding what a document or object is 'about'. What is actually done by Google or Bing is far more complex and has far more parameters at play, but the principles they employ of using statistical analysis to replicate human reasoning are similar to these simple methods of word frequencies. They are clearly amazingly successful: every Google search may bring up millions of 'hits' but it has a striking knack of highlighting the most relevant with remarkable consistency.

So is that the end of metadata? There are many good reasons why we might think so. Cost is an obvious but compelling one: there is no way we could pay people to read and abstract everything on the Internet in the way Google does with its secretive algorithms. Without content-based retrieval we would have to accept that we would not be able to find the vast majority of the riches that live out there, nor tell them apart from the dross. The Internet would be a much smaller place if we did not have Google and its peers.

Some might also argue that avoiding the human element, using statistics to weigh up the meanings of what appears in digital objects, is in some way purer, more unbiased, freer of the ideological underpinnings that inevitably find their way into all metadata because of its origins in humans and their thought. This is an appealing idea but a utopian one. The algorithms that Google uses are human constructs and the results they present are far from unfiltered representations of reality.

The parameters it employs to rank results and the decisions behind them are made by human beings and inevitably reflect their interests and preoccupations.

Google's algorithms and those of its rivals remain a secret but many people earn good money trying to second-guess where their emphases lie and so manipulate them for commercial advantage. A whole industry, known as Search Engine Optimisation (SEO), has grown up over the last 20 years that advises on making changes to the content of websites to push them up the rankings in search results. This is a very human world of competition and commerce, far removed from any ideals of semantic purity.

If we're looking to cast off bias from our journey through the Internet, abandoning metadata won't help us. But there are many other reasons why human beings and their metadata, imperfect as they are, have a role to play as gatekeepers to our collective knowledge. Using content-based retrieval to work out what something is 'about' is easier to do with some types of object than others. Text is relatively amenable to this and most of these techniques concentrate on it. But what about music or a painting? There are methods available for deciphering both of these but they are more basic than those developed for text. A human being can describe these and many others with much greater precision than any algorithm.

Metadata is also essential for context. We build knowledge and our conception of the world by placing its elements into relationship with each other. Even when we retrieve something on Google, we apply our very human sense of context to make sense of it. In the case of much of what we find on the Internet, our own interpretation may be all we need. But for plenty of material, it helps us greatly to have the knowledge and experience of others to guide us on our way to discovery and comprehension. It is in the areas where our personal knowledge meets its limits, and our ability to add this context is limited, that human-created metadata is important.

This is why the flowering of standards that we saw in Chap. 3 seems to be continuing unabated. The growth of the digital seems to need more metadata not less. Google and its peers make it possible to discover new material in ways which we could not have conceived of before but they need to be complemented by human thought and the metadata by which it is focussed. Metadata is far from dead but its silos need breaking down. One way of doing this, democratizing its production, forms the subject of the next chapter.

References

1. Noyes, D. (2013). *Line-mode browser dev days at CERN.* Restoring the first website. http://first-website.web.cern.ch/blog/line-mode-browser-dev-days-cern. Accessed 17 Nov 2015.
2. Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. *Scientific American, 5*, 29–37.
3. DBpedia, (2015). DBpedia. http://wiki.dbpedia.org/. Accessed 19 Nov 2015.
4. Shadbolt, N., Hall, W., & Berners-Lee, T. (2006). The Semantic Web revisited. *IEEE Intelligent Systems, 21*, 96–101.
5. Sikos, L. (2015). *Mastering structured data on the Semantic Web: From HTML5 microdata to linked open data*. New York: Apress.

Chapter 9
Democratizing Metadata

For most of its long and illustrious history, metadata has tended to be the preserve of the professional, the librarian, the archivist or the scholar. Although often intended for mass consumption, or at least wide circulation in specific communities, its creation has been very much the job of a small coterie of individuals. Some of these, such as Kallimachos of Cyrene, have been autodidacts who have taken it upon themselves to extend the practices that were current in their time into new areas; others (the vast majority) were either trained professionals or academics who produced it in the course of their research. Rarely has metadata been brought into existence by those who need to use it, those who lie outside these narrow circles of experts. It has been a top-down affair for centuries.

Many of the reasons for this were cultural or societal. For much of our history, education, even literacy, has been limited to narrow groups of people and the skills needed to create metadata to even smaller sections of the educated body. Part of the reason also lies in the technology used to create and house it. The handwritten catalogues of early libraries, their printed successors and the card catalogues that held sway until the technological changes of the 1960s did not lend themselves to contributions from outside their institutional or professional circles. Even during its first decades of automation, metadata remained centralized, the challenges of its abstruse formats requiring educated specialists capable of meeting their intellectual demands.

There have often been challenges to such elitist views of wisdom of any kind being the preserve of the few. Aristotle made the point in *Politics*, his grand work on government, that:-

> it is possible that the many, though not individually good men, yet when they come together may be better, not individually but collectively, than those who are so [1]

This notion has been encapsulated most widely in recent years in the pithy phrase "the wisdom of crowds". It formed the title of a popular book in 2005 by James Surowiecki [2] who asserted that bringing together large numbers of decisions individually reached could produce better results than those that emerged from the minds of so-called experts.

R. Gartner, *Metadata*, DOI 10.1007/978-3-319-40893-4_9

This is a rather romanticized notion and one that should be subject to caveat after caveat. There may not be much wisdom in crowds subject to collective hysteria or where they are not diverse enough to produce anything that a single individual could not come up with on their own. But in the area of metadata, the crowd may offer a rich source of knowledge, expertise or intelligence that could readily supplement or possibly replace the work of professionals.

To make the wisdom of crowds work needs some approach to allowing people to share their ideas and aggregate them into something more than the sum of their individual parts. Just when it was needed, along came a new label (if not a wholly original concept) that met this requirement: *Web 2.0.* This much vaunted term describes an approach to the Internet that gained popularity around the time of Surowiecki's book, the notion that its flow of information should no longer be one-way, from provider to consumer, but that everyone should be able to feed into it. The most dramatic manifestation of this philosophy has been the growth of social media services such as Facebook and Twitter. User-generated content is the key buzzword here: no longer are the keys to the Internet held by professionals alone, everyone can participate. It did not take long before some realized that metadata could benefit from the technologies and open outlook of Web 2.0 and that a more democratic model for its creation could deliver real benefits.

Social Cataloguing

Although Web 2.0 is a term that has gained traction in the twenty-first century, collaborative approaches to creating metadata have been around from well before it became a ubiquitous buzzword. A pioneering example still very much with us is the *Internet Movie Database* [3]. This started as an enthusiast's project in 1990 when a computer programmer named Colin Needham started publishing lists of film actors and actresses on forums in *Usenet,* a popular online discussion list service at the time. In 1993 he moved these to a separate website to which volunteers were asked to contribute information, hopefully carefully checked, on films old and new.

This has grown to be a huge success and for many film fans the primary port-of-call for everything related to cinema. Much of its content remains generated by its users, although this is now carefully checked by the service's own staff. It is certainly more widely used and more pervasive than any online reference source in cinema produced by professional or scholarly bodies. Similar stories abound in other areas where enthusiastic fans have been able to share their knowledge and expertise with their peers by cataloguing the objects of their interest. These include book cataloguing by bibliophiles (using such services as *LibraryThing* [4]), discographies by music fans, collections of recipes by foodies and bibliographies of comics by aficionados of that art form.

All of these projects are arguably as much about building communities as creating metadata. They have worked so effectively, and have produced surprisingly good results, because they tap into the knowledge of those steeped in their respective

sub-cultures who have become experts to some degree in their chosen speciality. Other projects have sought to spread their net more widely and harness the power of crowds to help them generate the type of data and metadata that would previously have been the preserve of highly-trained professionals.

Citizen Science from Galaxies to Ships' Logs

One of the things that human beings excel at is the recognition of patterns: huge amounts of computing power must go into emulating the dexterity at deciphering a visual pattern that a human can muster in a fraction of a second. Recognizing patterns is important in some areas of science and finding enough people to do this work can be beyond the budget of many a project. This was the problem that a group of astronomers faced in 2007 when confronted with something of a tidal wave of data. The *Sloan Digital Sky Survey,* a survey of galaxies and other distant objects begun in 2000 by the Apache Point Observatory in New Mexico, was producing images of galaxies at a pace well beyond the capacity of its team to process.

Two astrophysicists based at Oxford University, Kevin Schawinski and Chris Lintott, came up with the idea of using the Internet to harness the potential of millions of individuals whose interest in astronomy might tempt them to lend a hand in their task. So was born *Galaxy Zoo* [5], one of the first and most successful of all citizen science projects (Fig. 9.1).

Those entering the site are shown a picture of a galaxy taken by the project's 2.5 m optical telescope and asked a small number of simple questions about its shape and appearance. It takes about a minute to classify a galaxy in this way even for someone who is new to the process and has had no training in it. One of the draws of the project is that the image itself has probably never been seen before by another human being; it would have been taken by a robotic camera and processed by computer alone to reach the website and the eyes of its classifier [6].

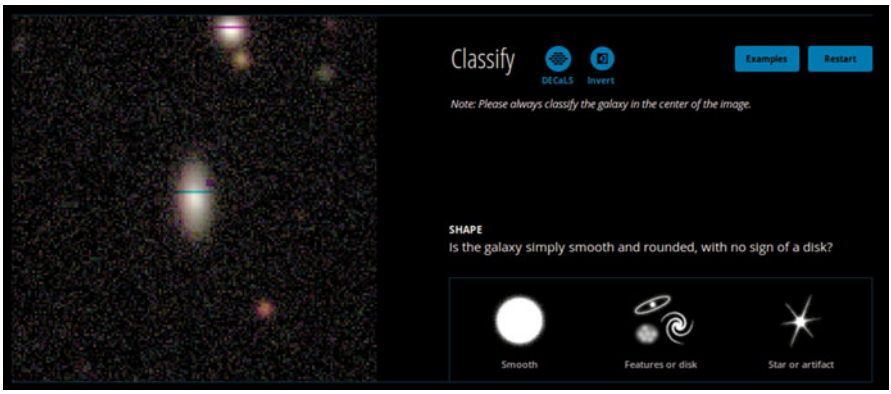


Fig. 9.1 Classifying a galaxy in Galaxy Zoo (galaxyzoo.org – a Zooniverse project)

The results were staggeringly successful. Over 100,000 people participated in the first trial, making 40 million classifications in its first 175 days of operation [7]. The same results would have taken several years for a single trained scientist to achieve working 24 hours a day [8]. Over the years the project has gone through several iterations, recruiting hundreds of thousands of further volunteers and clocking up hundreds of millions of classifications.

One big question mark hung over this project when it was first conceived, the quality and accuracy of the classifications themselves. Some scientists were highly sceptical that, in the words of one, "anybody off the street could come and do something better" than a trained researcher [9]. The way that the project gets round the vagaries of the untrained individual is to ensure that each galaxy is classified multiple times: in the first tranche, the average was 38 each [7]. Doing this pulls out inconsistencies very effectively, producing very high accuracy rates. Very few now question the validity of the results or the research based on them; they have appeared in multiple peer-reviewed articles in which their accuracy has been fully accepted [9].

Galaxy Zoo was an enormous success and many other projects followed in its wake. Astronomy proved a fertile area for citizen science and this pioneering project was followed by others which used crowd sourcing to classify craters on the Moon (Moon Zoo) [10], identify 'bubbles' in the interstellar spaces of the Milky Way (The Milky Way Project) [11] or hunt for planets outside the Solar System (Planet Hunters) [12]. Others have looked at the natural world, studying a range of animal species from whales to plankton.

Initially called citizen science, projects of this type began more regularly to be called crowdsourcing as many moved away from the purely scientific to embrace the humanities and social sciences. As they have expanded their remit, they have proved particularly adept at rescuing historical records and turning them into data capable of machine-readable analysis. One stand-out here is *Old Weather* [13], a project to reconstruct historical weather patterns from hand-written ships' logs of the late nineteenth and early twentieth centuries. Volunteers were asked to transcribe these logs into database tables, seemingly mundane and often tedious work but something that many took to with gusto. Transcribers could work on the logs for a given ship and become part of its 'crew' moving up the ranks as far as captain depending on the number of records they produced. The results were again outstanding and the data gathered has produced reliable pictures of weather patterns from a century ago.

The same techniques have proved usable in more traditional humanities fields. One groundbreaking project, *Ancient Lives* [14], asked volunteers to transcribe Greek papyri dating from the first to the sixth centuries CE. A knowledge of Greek was not required to do this as the site allowed the transcriptions to be made by pattern matching alone. The results were again remarkably successful. More conventional metadata has been produced by a project at the Bodleian Library in Oxford [15], home to some of the world's pioneering cataloguing projects in the seventeenth century. Here volunteers are asked to catalogue nineteenth-century musical scores that have already been digitized from the Library's collections. This project

has demonstrated that crowd-sourcing can produce traditional cataloguing metadata of some complexity.

The most successful of these projects are those that pay careful attention to motivating and retaining their army of volunteers. Some manage this by attractive websites which make the tasks engaging in themselves, some by the application of subtle incentives (such as the rankings on the Old Weather ships) and others by expending time and energy on maintaining discussion forums to engage their participants. They are by no means easy options to acquire cheap data and metadata with little effort, but when done effectively they can produce remarkable results which would otherwise have been impossible.

Folksonomy: Democratic Classification

For most of its history, one area of metadata that has almost exclusively been the domain of professionals or scholarly experts is classification, the process of deciding what something is 'about' and devising subject labels to express this. Deciding on the subject of something inevitably involves more of an interpretative act than documenting its title or author and this has often been entrusted to those considered experts in their domain. It has usually been assumed that they know more about the objects of their labours than those outside their august circles. But, as with the creation of metadata in crowdsoutcing, there is the potential here to widen the scope of classification to accommodate views that stretch beyond these narrow boundaries. To do this, we have to move from the venerable tradition of *taxonomy* to its more democratic modern counterpart *folksonomy.*

This word was first coined in 2007 by Thomas Vander Val [16], a consultant and information architect who came up with it to describe the classifications that appeared on such websites as the photo sharing platform *Flickr* [17]. These and similar services allow users to apply subject tags to the objects they have uploaded. There is no attempt to dictate or control what tags can be used, although most sites will often suggest some possible terms to make the process easier (in the case of Flickr by using shape recognition to get a rough idea of the subject of a photo). This is a social way of building up a classification from the bottom up, far removed from the rigid hierarchical model of traditional taxonomy.

Vander Val emphasized when he coined the term that folksonomies result from the *personal* tagging of objects; they are not intended to be of use to anyone except the individual doing the tagging. Each tag is a way for them to attach some meaning to an object so that they can retrieve it in the future and also to link it to others that they see as having the same meaning. But because these sites are social phenomena this conglomeration of personal subject tagging immediately takes on a collaborative and collective dimension. When combined with the tags created by others, each expressing in some way an interpretation of how their creators understand an object, something approaching the function of a standard taxonomy can be built up, although one very different in form and construction.

Fig. 9.2 A small folksonomy of Web 2.0 expressed in a tag cloud (Markus Angermeier and Luca Cremonini)

Many advantages have been claimed for this way of constructing classifications as compared to traditional taxonomy. One is its democratic appeal: the terms in a folksonomy come directly from the people who create and curate the objects and are expressed in the language they use and understand. This should in theory provide a much richer and potentially more multi-cultural vocabulary than classifications devised by committees of professionals. It should also encourage linguistic diversity, allowing tags in all languages to exist alongside those (particularly English) that predominate in most published taxonomies.

One potent visual image that has become popular as a way of expressing the shape of a folksonomy is the tag cloud. This displays its subject terms using font sizes to represent their relative importance. The cloud in Fig. 9.2 expresses a small set of terms associated with the Web 2.0 phenomenon.

The tag cloud has some use beyond being a pretty picture: it offers a way to filter the terms in a folksonomy by giving each a weighting, albeit in most cases by a relatively crude measure such as its frequency. Tag clouds do have their limits: it is difficult to scale one up beyond some tens of terms, after which it is likely to be too densely-packed to be usable. Other more standard ways of browsing a classification must be deployed after that.

Because they are constructed in the community and continuously updated as new objects are classified and new tags added, folksonomies easily outdo their traditional

forbears in terms of currency: they are inevitably more up-to-date with current thinking than formal schemes. A new term can emerge or even be invented by a single individual and instantly take its place in a folksonomy. As ideas change so can the tags that populate these classifications; there need be no time lag to allow them to gain widespread acceptance before they become 'legitimate'. For this reason they can reflect current thinking more accurately that their taxonomic cousins.

Because a single tag can be enough to establish a concept in a folksonomy, they can be particularly effective at describing what is often called 'long tail' content, the little known, obscure material that a standard taxonomy would not consider important or significant enough to bother with. A folksonomy can ensure that a minority interest does not disappear into a quagmire of popular or populist content, that it receives the recognition it deserves whether it is the preserve of a single individual or of millions. The proponents of folksonomy often emphasize its credentials as a builder of communities, but it has equal potential as a home for the lone voice.

These are all valid points although some are expressed with a degree of idealism that should set sceptical alarm bells ringing. The great strength of folksonomy is often claimed to be that it has a degree of authority because it comes directly from the people and presents an unfiltered representation of their living culture free of ideology. An appealing idea, but, as has been made clear in earlier chapters, the notion of metadata being devoid of ideology is a utopian one. Folksonomies are as ideological as any other form of metadata and what they present are beliefs about the world that are as value-laden as beliefs always are.

False idealism apart, there are other problems with this democratic mode of subject classification. Controlled vocabularies exist to bring clarity to a haze of terms that may describe the same concept; they do this by putting some shape and order into its synonyms, homonyms and alternative spellings. The free-for-all of folksonomy abandons attempts to do this, so there will inevitably be multiple ways of talking about the same thing. This is certainly democratic but it does mean low rates of retrieval: searching using a given term will inevitably mean missing records that describe the same concept but are tagged with an alternative one. Without some way of handling these thorny issues, we have to accept that we will miss plenty of material that could be relevant to us.

These need not be insurmountable problems: modern information science can come up with clever ways of alleviating them to improve search results. Some studies have shown that a consensus of terms rapidly takes shape when a folksonomy reaches a significant size [18]; this means that some sort of control emerges naturally. Sometime mathematics can be applied to folksonomies to iron out differences and map the idiosyncratic tags of individuals to terms that are recognized and used more generally [19]. A burgeoning area of research, called by some of its practitioners *folksontology* [20] is looking at how to generate and use the hidden structures of folksonomies in this way; this is already coming up with exciting results.

Folksonomy is certainly here to stay and is growing in popularity with every new move forward into the collaborative world of Web 2.0. It is certainly a challenge to traditional notions of taxonomy and the ways in which it is created. But it is unlikely to supersede its long-established precursor entirely for many of the reasons already

given in this book. What is needed is a *rapprochement* between the two, some way in which the best of both can be brought together.

Enrich Then Filter: Making Sense of the Metadata Morass

The emergence of user-generated metadata has produced an extraordinarily rich but very messy resource. Pretty well everything that the human mind is capable of conceiving can be found there somewhere, some of it readily comprehensible to everyone, some of it the preserve of its creators or a small circle of their friends and admirers. It has produced more metadata that is more current and in many ways more culturally vibrant than the traditional methods that have appeared in most of this book. The challenge is how to make sense of it, how to tap into this resource without getting buried in its quagmire.

The answer seems to be working out some way of filtering it, honing it down so that it becomes usable and lets us find what we want. This is the model proposed in a recent book by Getaneh Alemu and Brett Stevens; they use the epithet 'enrich then filter' for a model that should let us have the best of both worlds, diverse, vibrant metadata and focussed, relevant search results [21]. Their idea is that we encourage metadata from all sources: the 'expert'-created records that have been the norm for so long and community-generated ones that come from anyone willing to provide it. We then filter this rich but confusing and always changing body of metadata as and when we need it to meet our requirements.

Filtering, of course, is done whenever we use metadata to find something we want. Everyone who has done a search, employing anything from the hand-carved lists of ancient Babylonia to the modern universal database of the Semantic Web, has had to filter what they come up with to make sense of it. Their task is generally easier if they search records that have been compiled with some consistency. It is much simpler to find what we need from a carefully structured MARC record or a rigidly defined ontology than a mass of folksonomy tags. The challenge is to make this possible for the unstructured mess that enriched metadata can rapidly become.

An important feature of this must be flexibility. The results we want will differ according to who we are, what we are interested in, whether we want something precise or are browsing serendipitously and so on. The idea is to move away from a monolithic view of the users of metadata to one that recognizes their diversity. Enriched metadata should provide ample raw material to embrace this: what is needed are ways to serve up different views of it that are tailored to the idiosyncratic needs of those at the end of the search chain.

This is easier said than done but it is undoubtedly the way forward. Metadata has undergone a fundamental change since the Internet and particularly the two-way channels of Web 2.0 have come along; the forces behind its democratization have now become irresistible. It is now hard to imagine many a well-known service such as Amazon operating without user-created contributions nor should we want to. Equally there is no reason why we should throw out the hard work of metadata

experts who have honed their techniques over thousands of years. The challenge is to marry the two to produce something more powerful than either can provide on its own. 'Enrich then filter' seems the right model to adopt. Its implementation will need hard work by theorists and developers to put these ideas into practice, but the rewards will certainly be worth it.

One of the most over-used phrases in much writing, and one of which information scientists are as guilty as anyone, is 'paradigm shift', a term devised by the physicist Thomas Kuhn to describe the sudden fits and starts by which science suddenly realigns itself when existing models reveal themselves to be inadequate [22]. Almost anything can be described as a paradigm shift to give it spurious importance, so I will not use it to describe the advent of democratic metadata. It does, however, represent one of the most significant challenges in generations to the ways in which we need to look upon our 'data about data'. Metadata, as always throughout its history, never stands still.

References

1. Aristotle. (1944). *Aristotle in 23 volumes*. Cambridge, MA: Harvard University Press.
2. Surowiecki, J. (2005). *The wisdom of crowds*. New York: Anchor Books.
3. Internet Movie Database. (2015). IMDb – Movies, TV and Celebrities – IMDb. http://www.imdb.com/. Accessed 11 Dec 2015.
4. LibraryThing. (2015). LibraryThing | Catalog your books online. https://www.librarything.com/. Accessed 11 Dec 2015.
5. Zooniverse. (2015). Galaxy Zoo. http://www.galaxyzoo.org/. Accessed 7 Dec 2015.
6. Hopkin, M. (2007). See new galaxies – Without leaving your chair. *Nature*. http://www.nature.com/news/2007/070709/full/news070709-7.html. Accessed 7 Dec 2015.
7. Lintott, C., Schawinski, K., & Bamford, S. (2011). Galaxy Zoo 1: Data release of morphological classifications for nearly 900 000 galaxies. *Monthly Notices of the Royal Astronomical Society, 410*, 166–178.
8. Pinkowski, J. (2010). How to classify a million galaxies in three weeks. *Time*. http://content.time.com/time/health/article/0,8599,1975296,00.html. Accessed 26 Feb 2016.
9. Secorum Palet, L. (2014) Crowdsourcing science goes boom. *USA Today*. http://www.usatoday.com/story/news/nation/2014/07/25/ozy-crowdsourcing-science/13143465/. Accessed 8 Dec 2015.
10. Zooniverse. (2015). Moon Zoo. http://www.moonzoo.org/. Accessed 8 Dec 2015.
11. Zooniverse. (2015). The Milky Way Project. http://www.milkywayproject.org/. Accessed 8 Dec 2015.
12. Zooniverse. (2015). Planet Hunters. http://www.planethunters.org/. Accessed 8 Dec 2015.
13. Zooniverse. (2015). Old Weather. http://www.oldweather.org/. Accessed 8 Dec 2015.
14. Zooniverse. (2015). Ancient Lives | Help us to Transcribe Papyri. http://www.ancientlives.org/. Accessed 8 Dec 2015.
15. Zooniverse. (2015). What's the score at the Bodleian. http://www.whats-the-score.org/. Accessed 8 Dec 2015.
16. Vander Wal, T. (2007). Folksonomy :: vanderwal.net. http://www.vanderwal.net/folksonomy.html. Accessed 10 Dec 2015.
17. Flickr. (2015). Flickr, a Yahoo company | Flickr – Photo Sharing! https://www.flickr.com/. Accessed 10 Dec 2015.

18. Robu, V., Halpin, H., & Shepherd, H. (2009). Emergence of consensus and shared vocabularies in collaborative tagging systems. *ACM Transactions on the Web (TWEB), 3*, 14.
19. Wetzker, R., Zimmermann, C., Bauckhage, C., & Albayrak, S. (2010). I tag, you tag: Translating tags for advanced user models. In *Proceeding of the third ACM international conference on web search data mining* (pp. 71–80). New York: ACM.
20. Van Damme, C., Hepp, M., & Siorpaes, K. (2007). Folksontology: An integrated approach for turning folksonomies into ontologies. In *International workshop Bridging the Gap between Semantic Web and Web 2.0 (SemNet 2007) 4th European Semantic Web conference* (pp. 57–70), Innsbruck, Austria.
21. Alemu, G., & Stevens, B. (2015). *An emergent theory of digital library metadata: Enrich then filter*. Waltham: Chandos Publishing.
22. Kuhn, T. S. (1970). *The structure of scientific revolutions*. Chicago: University of Chicago Press.

Chapter 10
Knowledge and Uncertainty

One of the most memorable moments in television history occurred in an episode of the 1970s BBC series *The Ascent of Man*. This remarkable series attempted to tell the history of human society through the development of our scientific understanding. In one episode, entitled *Knowledge or Certainty*, its presenter, the Polish-born biologist and historian-of-science Jacob Bronowski, stands in a pond within the compound of the Auschwitz concentration camp, the resting place for the ashes of millions murdered there. Holding some of the mud from the pond in his hands, he issues a grave warning against those who claim with complete certainty to know everything. "When people believe that they have absolute knowledge", he says, "with no test in reality, this is how they behave. This is what men do when they aspire to the knowledge of gods" [1].

Bronowski points out with great force the arrogance of claiming omniscience and its potentially dire consequences. He shows that certainty is impossible in itself: there is, he claims "no absolute knowledge…all information is imperfect. We have to treat it with humility" [1].

Other great minds before him had examined the limits to understanding and shown that not only is such an assumption of complete knowledge dangerous, it is logically impossible. They include the philosophers Immanuel Kant, Artur Schopenhauer and Ludwig Wittgenstein who have effectively buried forever the notion of absolute knowledge. Science has also shown us the limits of knowledge to humbling effect. From Schroedinger's unfortunate cat caught between life and death at the whim of a decaying particle to Heisenberg's uncertainty principle nature proclaims our ignorance with compelling force. Science is rather more humble in its claims to knowledge than it has been in centuries past because uncertainty appears to be written into the fabric of the universe.

As we have seen throughout this book, metadata has an essential but often invisible role in the way we build our knowledge. It allows us to bring together its atoms, the single units of data into which we can divide the world, into the information that gives them meaning. It then lets us assemble this information into knowledge. Each of these moves up the ladder, or up Ackoff's pyramid which we met in Chap. 1, is

R. Gartner, *Metadata*, DOI 10.1007/978-3-319-40893-4_10

achieved by joining the bits and pieces that make up each level and putting them into context. When we have done this, we can find patterns of meaning to help us fulfil the very human need to interpret what we have and make sense of it. Metadata lets us forge these links and construct our edifices of knowledge from the smallest bricks of data. The clue to how this is done is in our basic, etymological definition of metadata that was introduced at the beginning of this book: 'data about data'. 'About' is what metadata does: human knowledge is built on 'aboutness' and it is through our interpretation of what the world is 'about' that most of our intellectual endeavours are based. Without metadata we cannot have knowledge, at least according to the epistemological model, influenced by Ackoff, that we have followed in this book.

Of course, other epistemologies are available and one that concentrates solely on metadata and its role would be so restrictive as to deny much of what is interesting about humans and their intellects. This is a mechanistic view of knowledge, one which is undoubtedly accurate in its essence but perhaps unrevealing. It is tempting to say, 'yes, this is true, but so what?'.

One 'so-what' that this model can make us aware of is the need for modesty. Bronowski highlighted dramatically the dangers of claims to omniscience as he stood in the ashes of the victims of Auschwitz. A view of knowledge founded on metadata should rapidly debunk any notions of it being absolute. It should be obvious that metadata has its limits: it is an interpretation of the world, an abstraction that throws away as much as it keeps. Few who look at it in any depth could consider that it contains unassailable truths, that it is much more than the codified expression of how some view the universe. It is inherently ideological as we have seen. For all of the apparent 'science' behind it, it is fundamentally an expression of world views and no more than that. It is hard to see a system of knowledge built on these foundations ever telling us anything and everything with the arrogance of perfection. Seeing the role of metadata in building knowledge should allow us to heed Bronowski's call to treat knowledge and the information that makes it up with the humility he calls for.

To emphasize the role of metadata in encouraging modesty is not to underestimate its importance. We have seen that little of human intellectual effort could have been achieved or passed on without it. Its role in the creation of culture is difficult to ignore. Museums, those inspiring edifices which allow the intellectual achievements of one generation to be passed to the next are physical manifestations of the ways in which curation, the 'caring-for' of the products of our minds, is key to the building of culture. Museums cannot exist without metadata, nor can the cultures that they represent. Princess Ennigaldi, the Babylonian princess whose embryonic museum so excited Leonard Wooley in the 1920s, understood this when she catalogued and described the artefacts she had so carefully collected. Her cylindrical labels, so painstakingly constructed, showed that documenting her finds and putting them into context was as an important part of their curation as the physical handling of their storage and preservation.

If metadata is key to museums, it is even more central to libraries, those equally ennobling edifices which allow human knowledge to engage with its past and

propagate in its present. Not for nothing have librarians figured so prominently in the narrative of metadata. The work of pioneers such as Alexandria's Kallimachos is still inspiring in its intellectual daring and rigour. Attempting to shape the amorphous mass that human knowledge may offer us can be a daunting prospect and it takes some degree of inner conviction to take it on. Other innovators whose names are now largely forgotten, including Thomas James of the Bodleian and Henriette Avram, the inventor of MARC, have had more influence on the development of human intellectual achievement than many a better-known scholar. Modesty may be a key part of metadata, and a refreshing one, but perhaps some of its greatest minds deserve more recognition in the history of scholarship.

As we have entered the age of the Internet, the ingenuity of those who push the bounds of metadata has not diminished. A grandiose vision such as Tim Berners-Lee's Semantic Web can still inspire us and propel us to see new ways at looking at how we relate to information and knowledge. Practicalities may sometimes clip our wings when we try to get these ideas going but they are elevating enough to motivate us to continue putting our efforts into making them work.

In a different way the growth of crowdsourcing as a way of creating metadata also says much of human ingenuity and its altruistic side. So many people spending so many hours helping to push forward knowledge by describing the world, looking for patterns in it and sharing what they find, even when the work required is mundane and repetitive, testifies to the inspiration that knowledge for knowledge's sake remains even in a world of neoliberal market ideologies. That the creation of metadata is seen by those giving their time in this way as a valid way to move knowledge forward is a sign of its centrality to human thought. It remains as ever a potent gateway which, by putting shape to the amorphous cloud of knowledge, allows us to make sense of it. By generating metadata we assert our desire to understand and grow our understanding.

Metadata, like knowledge, has never stood still. We might expect this if we take as read our most basic definition for it, 'data about data': data and its higher aggregations, information, knowledge, understanding and maybe even wisdom, have never stopped changing and so inevitably has metadata. But history shows that its development has not just been a passive reflection of the changes in what it describes. The pioneers of metadata have made their own contributions to the way in which we understand the world through the philosophical underpinnings of their work. The hierarchies of a taxonomy, the network of semantic links of an ontology, the seething mass of semantic molecules that makes up the Semantic Web, all of these tell us something of their creators' models of understanding and knowledge. They are epistemological statements, all theoretically reflecting the same 'reality' but expressing worldviews as divergent as many a philosophical disagreement often is.

Which brings us back to ideology, not necessarily in a derogatory sense, nor even in the sense of the definition adopted in much of this volume as a set of beliefs presenting a supposedly transparent reflection of reality. Instead, it could be useful to invoke the first meaning of the word given it by Antoine Destutt de Tracy, ideology as a 'science of ideas'. Metadata is in many ways an attempt to develop a science for organizing ideas and so creating knowledge. We can follow a model such as Ackoff's

pyramid to visualize this but we need not resort to such formal ways to appreciate what it attempts to do. If we do go back to etymology, we might consider metadata one of the most ideological of disciplines, but in a positive sense. By attempting to make the organization of knowledge more scientific and rational we allow it to be constructed, to advance and to develop.

Back to Bronowski and *The Ascent of Man*. It is difficult to imagine much of the 'ascent' he describes ever gaining purchase without some sort of metadata to guide us along each step of the way. Scientific knowledge has always relied on the work of previous generations, the 'shoulders of giants' as Isaac Newton described it [2], to push its boundaries. Metadata has always been key to allowing knowledge to be curated and transmitted between generations. Cities could not have been built without it nor would the developments of the industrial revolution, which have given us the world we experience today, have been possible if it did not exist.

One of the most telling symbols of the human need to abstract the world, the practice on which so much of the 'science' of metadata is built, come from the end of the first episode of Bronowski's series. Here he visits the caves at Altamira in Spain, famous for their palaeolithic animal paintings. These are basic but enthralling metadata, saying something about the world they describe and their creators. Bronowski argues that they are a way in which humans can look to the future, plan for it and prepare for its challenges. They curate their knowledge of the time, particularly those skills at hunting on which their survival depended, and put it in concrete form so that others may understand what lies ahead for them.

Even more resonant to Bronowski, and to us, is a drawing of a human hand. The 'data', the human on whom it was modelled, is long gone, the 'metadata', this abstraction of a part of their body, remains. The hand, says Bronowski, tells us "This is my mark. This is man" [1]. Metadata is how we make our mark on the world by forging our knowledge. It is a cold technical term but there is really nothing cold or technical about it, it is part of the human condition. It represents our desire to make sense of the world, to know what it, and we, are 'about'.

References

1. Bronowski, J. (1981). *The ascent of man*. London: Futura.
2. Newton, I. (1959). *The correspondence of Isaac Newton/1: 1661–1675*. Cambridge: Cambridge University Press.

Index

R. Gartner, *Metadata*, DOI 10.1007/978-3-319-40893-4