CHAPTER 4

Basic concepts of information science

Where is the life we have lost in living?
Where is the wisdom we have lost in knowledge?
Where is the knowledge we have lost in information?
T. S. Eliot, *Choruses from 'The Rock'*

It is hardly to be expected that a single concept of information would satisfactorily account for the numerous possible applications of this general field.
Claude Shannon

Information is information, not matter or energy. Norbert Wiener

Introduction

In this chapter, we will consider some of the basic concepts of the information sciences: information and knowledge, documents and collections, relevance and 'aboutness', and information use and users.

It may seem strange to find that there is still debate about the nature of these very fundamental ideas: as strange, perhaps, as finding a doctor who had no idea what a 'disease' or a 'treatment' was, or an engineer who had no idea what was meant by 'materials' or 'design'. That is not to say that there need be a perfect understanding of these concepts; doctors treated diseases, sometimes quite effectively, long before they had any realistic idea of what caused them. But most professions expect to have some understanding of the basic concepts with which they deal.

'Information', 'knowledge', 'document', and so on, are tricky concepts, which can have many different meanings, and can be understood in many different ways. These are not just academic matters; they can have a real effect on professional practice. What someone understands by 'knowledge', for example, and its relation to 'information', will determine how they go about the practical business of 'knowledge management'. And what a librarian or information specialist understands by a 'document' will determine what sort of things they keep on their shelves or in their computer files.

We begin by looking at perhaps the most fundamental of concepts: information itself, and knowledge.

Information and knowledge

Shannon and Wiener and I
Have found it confusing to try
[illegible]
To measure sagacity
And channel capacity
By $\Sigma\ p_i \log p_i$

Anonymous, *Behavioural Science*, 1962, 7 (July issue), 395

Information, argued John Feather and Paul Sturges in the 1997 Routledge *International Encyclopaedia of Information and Library Science*, is probably the most used, and the least precisely understood, term in the library and information world.

The best way to understand the concept of 'information' has been debated for many years; the more so as the ideas of the 'information age' and 'information society' have gained currency. The 'commonsense' meaning of the word relates to knowledge, news or intelligence, given and received, so that someone becomes 'informed'. But the word has had many different meanings over the years; its entry in the full Oxford English Dictionary of 2010, which shows its usage over time, runs to nearly 10,000 words.

Even within an information science context, the term has been understood with different connotations, as we shall see later.

To show the variety of ideas available, information has been explained, among many other things, as being as shown in the box below. This subset of the many definitions and explanations available shows their diversity, from simple to complex. Some relate information to concepts such as 'data' or 'knowledge', which then themselves require explanation. Some set the idea of information firmly in the context of communication between people; others see it as a more general concept, to do with structure, pattern, and organization. In this chapter, we can do no more than mention some of the more influential and interesting ideas put forward, allowing the reader to follow these up, and to find others, via the references.

Despite much effort, there is still no consensus as to what information 'really is', still less any 'theory of information', usefully applicable in all contexts; we will look at two theories for which this claim may be made later. The problem is made worse by the fact that the word is used rather differently in many different subject areas; for an early, and very influential, overview of this diversity, see Machlup and Mansfield (1983). As Capurro and Hjørland (2003, 356 and 396) say:

> . . . almost every scientific discipline uses the concept of information within its own context and with regard to specific phenomena . . . There are many concepts of information, and they are embedded in more or less explicit theoretical structures.

Varied conceptions of information: 'information is . . . '

fragmented knowledge (Bertie Brookes)

knowledge packaged for a user (CILIP Body of Professional Knowledge)

meaningful data (Luciano Floridi)

an assemblage of data in a comprehensible form capable of communication and use (John Feather and Paul Sturges)

patterns of self-organized complexity, providing meaning-in-context and promoting understanding (David Bawden)

communicated signs (Claude Shannon)

a change of structure (Nick Belkin and Steve Robertson)

a stimulus originating in one system that affects the interpretation by another system of either the second system's relationship to the first or of the relationship the two systems share with a given environment (Andrew Madden)

a difference which makes a difference (Gregory Bateson)

some pattern of organization of matter and energy given meaning by a living being (Marcia Bates)

an abstract concept, which manifests itself by organizing systems (Tom Stonier)

We will now look briefly at the use of the information concept in different contexts; specifically in the physical and biological sciences, as opposed to its more familiar usage in the social and information sciences. We can only consider this in outline, referring the reader to more detailed accounts below.

Information: physical, biological, social

The idea of information as a feature of the physical world arose through studies of the thermodynamic property known as *entropy*, through the work of physicists such as Ludwig Boltzmann and Leo Szilard. Entropy, usually understood as a measure of the disorder of a physical system, is also associated with the extent of our knowledge of it; put crudely, if a system is disordered, we have little knowledge of where its components are, or what they are doing. A formal mathematical link may be made between entropy and information, when information is defined in the way required by Shannon's theory, discussed below. Analysis of the relation between information and physical entropy led Rolf Landauer to propose his well known aphorism 'information is physical'.

Information must always be instantiated in some physical system; for the kinds of information of interest to the information sciences, this would be a human brain or some kind of document.

The idea of information as a physical entity has received increasing attention in recent decades. Information has been proposed as a fundamental aspect of the physical universe, on a par with – or even more fundamental than – matter and energy. One of the originators of this approach was the American physicist John Wheeler, who coined the phrase 'it from bit' to express the idea that physical reality is, in some way, generated from information; his views are surveyed, critiqued and extended in papers in Barrow, Davies and Harper (2004). Currently active proponents are Lee Smolin, who has suggested that the idea of space itself may be replaceable by a 'web of information', and David Deutsch, who proposes that information flow – essentially what changes occur in what order – determines the nature of everything that is. 'The physical world is a multiverse', writes Deutsch (2011, 304), 'and its structure is determined by how information flows in it. In many regions of the multiverse, information flows in quasi-autonomous streams called histories, one of which we call our universe'.

Similarly, in biology, the discovery of the genetic code and associated developments in molecular biology have led to the idea that information is a fundamental biological property, and that the transmission of information may be as fundamental – or more fundamental – a property of living things as metabolism, reproduction, and other signifiers of life.

Fascinating though these ideas are, we cannot pursue them further here; interested readers should begin with the overviews given by Gleick (2011), Floridi (2010), Davies and Gregersen (2010), Vedral (2010) and von Baeyer (2004).

The question then naturally arises as to whether these scientific ideas of information have any relevance for the information sciences, as we understand them in this book; or whether it just happens that the English word 'information' is used to mean quite different things in different contexts.

Very different views have been taken. Some authors have devised quite elaborate schemes to link physical, biological and human information. The British academic Tom Stonier, in a series of three books, advanced a model of information as an abstract force promoting organization in systems of all kinds: physical, biological, mental and social, and including recorded information (Stonier, 1990; 1992; 1997).

Marcia Bates, a well known American information science scholar now Professor Emerita at the University of California Los Angeles, has advanced a similar all-encompassing model, which she terms 'evolutionary' (Bates, 2005; 2006). It relies on a number of inter-related 'information-like' entities:

- Information 1 – the pattern of organization of matter and energy
- Information 2 – some pattern of organization of matter and energy given meaning by a living being
- Data 1 – that portion of the entire information environment available to a sensing organism that is taken in, or processed, by that organism
- Data 2 – information selected or generated by human beings for social purposes
- Knowledge – information given meaning and integrated with other contents of understanding.

This model, while all-encompassing and one of the more ambitious attempts at integrating information in all its contexts, remains at a conceptual and qualitative level, and introduces a potentially confusing multiplicity of forms of information and similar entities.

Others have denied that it is at all useful to try to make such a direct link. Cole (1994) and Hjørland (2007), for example, argue against any equating of the idea of information as an objective and measurable 'thing' to the kind of information of interest in library and information science; this information, they argue, is subjective in nature, having meaning to a person in a particular context.

Still others have proposed that it is an open, though interesting, question. One of the authors of this book has suggested that we may see information in human, biological and physical realms as being related through emergent properties in complex systems; one does not look for direct equivalences, but for more subtle linkages (Bawden, 2007a; 2007b).

It seems clear that the question as to whether there can be any generally applicable theory of information in all contexts is an important one. However, as Jonathan Furner (2010, 174) puts it 'the outlook for those who would hold out for a 'one size fits all' transdisciplinary definition of information is not promising'. The best known potential contender so far is Shannon's information theory, to which we now turn.

A mathematical theory of information, with a little semiotics

The closest approach to a universal formal account of information is Shannon's 'information theory', originated by Claude Shannon, and properly referred to as Shannon-Weaver-Hartley theory, in recognition of those who added to it and gave it its current form. This gives a rigorous mathematical basis for calculating the amount of information that can be transmitted through a medium or channel, and is an important tool for communication engineers. It has been widely applied to introduce the concept of information into the physical and biological sciences. But – despite much effort – this theory has made little impact on the theory or practice of librarianship or information science. As the limerick which opens this section reminds us, it is not easy to use a single equation to calculate the value both of telecommunications capacity and of human knowledge.

The first quantitative measures of information came from the US Bell Laboratories, which supported the work of the three pioneers – Harry Nyquist, Ralph Hartley and Claude Shannon – and in whose in-house journal the first writings in what would come to be termed 'information theory' were published. These focused, naturally enough, on the set of engineering issues around the transmission of messages across various kinds of physical communication network. Gleick (2011) gives a good account of their work.

The initial steps were taken by Nyquist, who in 1924 showed how to estimate the amount of information which could be transmitted in a channel of given bandwidth; the focus of his paper was on the telegraph. He did not, however, use the term 'information', writing instead of the 'transmission of intelligence'. These ideas were developed by Hartley in 1928, who established a quantitative measure of information, so as to compare the transmission capacities of different systems. He titled his paper 'Transmission of information', and used the word throughout. Nyquist and Hartley thereby arguably originated one of the modern uses of the term.

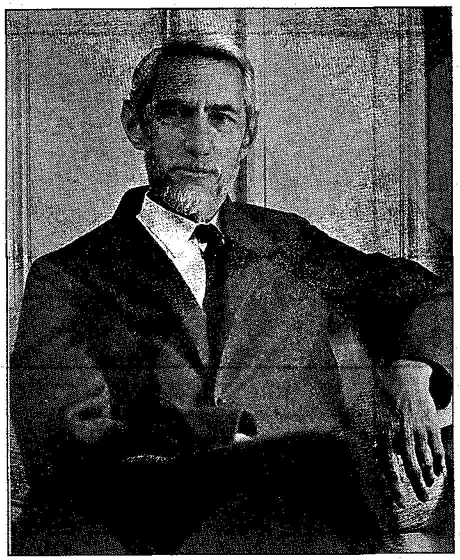


Figure 4.1
Claude Shannon (reproduced courtesy of the Library of Congress)

Hartley emphasized that this measure was 'based on physical as contrasted with psychological considerations'. The meaning of the messages was not to be considered; information was regarded as being communicated successfully when the receiver could distinguish between sets of symbols sent by the originator. His measure of information, understood in this way, was the logarithm of the number of possible symbol sequences. For a single selection, the associated information, H, is the logarithm of the number of symbols, s.

$$H = \log s$$

This in turn was generalized by Claude Shannon (Figure 4.1) into a fuller theory of communication, in an article of 1948, which was later republished in book form (Shannon and Weaver, 1949). It is interesting to note that the original modest claim of the article to be describing *A* mathematical theory of communication had been transmuted by the time the book was published to *The* mathematical theory. It was never claimed to be a theory of information, and it is perhaps unfortunate that it has been widely known as such; it is better to stick with Shannon's title of mathematical theory of communication (MTC). Its basic characteristics are set out in the box here.

Warren Weaver, a distinguished mathematician, scientific administrator and proponent of the public understanding of science, contributed an essay entitled *Recent contributions to the mathematical theory of communication* to the book. Roughly paralleling and expounding Shannon's more lengthy technical article, Weaver's contribution – philosophical, non-mathematical, wide-ranging, and considerably easier to read – has arguably had greater influence in spreading the ideas of information theory than any of its originators.

Shannon's Mathematical Theory of Communication

Shannon's article confines itself to a mathematical and technical derivation of what was still being referred to as communication, rather than information, theory. Shannon, noting that he is following Nyquist and Hartley in developing general theory of communication, defined the fundamental problem of communication as the accurate reproduction at one point of a message selected from another point. Meaning is ignored: 'these semantic aspects of communication are irrelevant to the engineering problem' (Shannon and Weaver, 1949, 3). What matters is that the actual message in each case is one which is selected from the set of possible messages, and the system must cope with any selection. If the number of possible messages is finite, then the information associated with any message is a function of the number of possible messages.

Shannon derives his well known formula for H, the measure of information

$$H = - K \sum p_i \log p_i$$

where p_i is the probability of each symbol, and K is a constant defining the units. The minus sign is included to make the quantity of information, H, positive; necessarily a probability will always be less than 1, and the log of such a number is always negative.

Shannon pointed out that formulae of the general form $H = - \sum p_i \log p_i$ appear very often in information theory, as measures of information, choice and uncertainty; the three concepts seem almost synonymous for his purposes. Shannon then gave the name 'entropy' to his quantity H, since the form of its equation was that of entropy as defined in thermodynamics.

Weaver took a much broader scope in his essay, generalizing Shannon's purely engineering concept of information. But he had to note that 'the concept of information developed in this theory at first seems disappointing and bizarre – disappointing because it has nothing to do with meaning, and bizarre because it deals not with a single message but rather with the statistical character of a whole ensemble of messages, bizarre also because in these statistical terms the two words *information* and *uncertainty* find themselves to be partners' (Shannon and Weaver, 1949, 116). Weaver, however, argued that these are merely temporary reactions, and that a 'real theory of meaning' might follow.

Shannon's was not the only attempt to derive a mathematical theory of information, based on ideas of probability and uncertainty. The British statistician R. A. Fisher derived such a measure, as did the American mathematician Norbert Wiener, the originator of cybernetics. The latter's mathematical formalism was the same as Shannon's but, confusingly, he treated information as the negative of physical entropy, associating it with structure and order, whereas Shannon equated information with entropy. Shannon's information is, in effect, the opposite of Wiener's, which has caused confusion ever since for those who seek to understand the meaning of the mathematics. A similar idea to Wiener's conception of information as negative entropy had been proposed by the German physicist Erwin Schrödinger, one of the pioneers of quantum mechanics, who had suggested that living organisms feed upon negative entropy. The idea of information as the opposite of entropy was popularized, with the snappy title of 'negentropy', by Leon Brillouin (1956), who first advocated the wide use of information theory in science. It has since been applied widely, some might say too widely, in subjects including economics, psychology and theology. For overviews of these intriguing ideas, see Floridi (2010) and Gleick (2011).

Shannon's theory gives the only convincing quantitative measure of information yet derived, and a great deal of work has been done to try to apply it in a variety of disciplines. Numerous attempts have been made to extend it to deal with meaningful semantic information, and to develop mathematical models for information flow, by authors such as Bar-Hillel, Dretske, Devlin and Barwise; for overviews, see Floridi (2011a) and Cornelius (2002), and for interesting examples see Karamuftuoglu (2009) and Cole (1997). However, these have had very limited impact on the theory and practice of information science.

The reason for this can be understood from a consideration of information from a semiotic viewpoint (see, for example, Liebenau and Backhouse, 1990). This shows us that any communication of information can be understood at four 'levels':

- empiric: the physical transmission
- syntactic: the language or coding used
- semantic: the meaning of the message
- pragmatic: the significance of the message to a recipient in a particular context.

'Information theories', following Shannon's MTC, deal almost exclusively with the syntactic level; while libraries and information services – though they must consider these levels – are generally more concerned with meaning and significance of information. Hence the failing of these theories, so far, to inform the principles and practice of librarianship and information science.

We will now turn to how information, as it is generally regarded within the sciences which focus on human recorded information, is understood.

Information for information science

Even when we consider only information within the library and information disciplines, there are a variety of views as to how information is to be understood; see Bawden (2001) for a short overview, Ma (2012), Cornelius (2002) and Capurro and Hjørland (2003) for detailed reviews, and Belkin and Robertson (1976) and Belkin (1978) for older, but still interesting, perspectives.

In an influential paper from 1991, Michael Buckland distinguished three uses of the term 'information':

- Information-as-thing, where the information is associated with a document
- Information-as-process, where the information is that which changes a person's knowledge state
- Information-as-knowledge, where the information is gauna equated with the knowledge which it doda' imparts.

Information-as-thing regards information as physical and objective; or at least being 'contained within' physical documents, and essentially equivalent to them. The other two meanings treat information as abstract and intangible. Buckland gives arguments in favour of the information-as-thing approach, as being very directly relevant to information science, in as much as it deals primarily with information in the form of documents; as we shall see later in this chapter, Buckland is associated with the 'documentation' approach to the subject.

Information-as-process underlies theories of information behaviour with focus on the experience of individuals, such as those of Dervin and Kuhlthau, which we shall consider in a Chapter 9.

Information-as-knowledge invokes the ideas that information and knowledge are closely related; as does the formulation of Bates, noted above. The exact relation, however, is not an obvious one.

The ideas, though useful, do not constitute a precise or formal description of information. We have seen that the information theoretic approach has severe limitations for our purposes. The only current candidate is the General Definition of Information (GDI), proposed by Luciano Floridi as part of his Philosophy of Information, discussed in Chapter 3 (Floridi, 2011b; 2010). As we noted there, Floridi analyses the ways in which information may be understood, and opts to regard it from the semantic viewpoint, as 'well formed, meaningful and truthful data'. Put formally, the GDI states that:

GDI) σ is an instance of information, understood as semantic content, if and only if:
GDI.1) σ consists of *ņ data*, for n ≥≥ 1;
GDI.2) the data are *well formed*;
GDI.3) the well formed data are *meaningful*.

Data is understood here as simply a lack of uniformity; a noticeable difference or distinction in something. To count as information, a collection of data must be well formed (put together correctly according to relevant syntax), meaningful (complying with relevant semantics), and truthful; the latter requires a detailed analysis of the nature of true information, as distinct from mis-information, pseudo-information and false information. A 'map' of the full set of types of information in this theory is shown in Figure 4.2; note the positioning of information in Shannon's sense away from issues of truth, meaning and knowing.

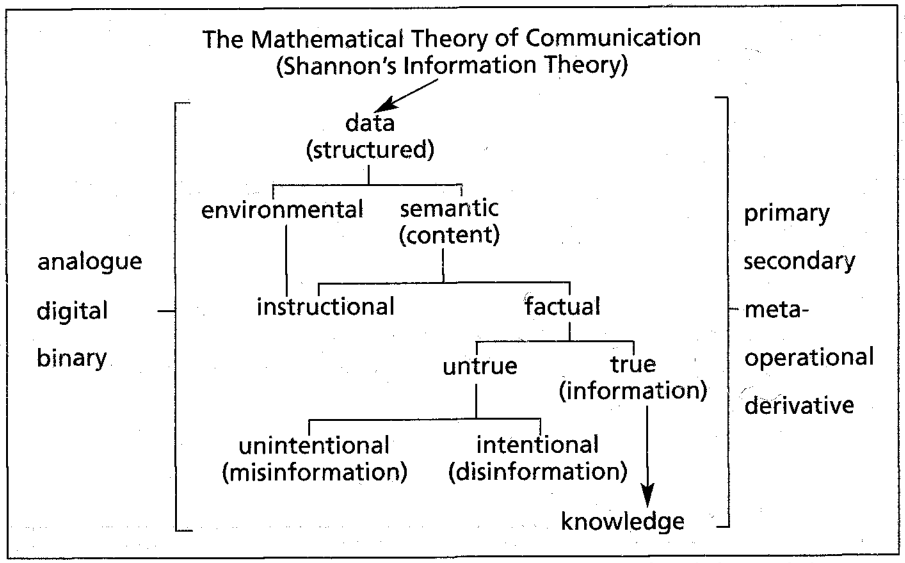


Figure 4.2 *Floridi's information map (reproduced from Floridi (2011a), by permission of Oxford University Press)*

Floridi sees information as related to knowledge, since knowledge and information are part of the same 'conceptual family'. Information is 'converted' to knowledge by being interrelated, which may be expressed through network theory. Informally:

> what [knowledge] enjoys and [information] lacks . . . is the web of mutual relations that allow one part of it to account for another. Shatter that, and you are left with a pile of truths or a random list of bits of information that cannot help to make sense of the reality that they seek to address.
>
> Floridi (2011b, 288)

Furthermore, information which is meaningful must also be relevant in order to qualify as knowledge, and this aspect may also be formally modelled, as also the distinction between 'knowing', 'believing' and 'being informed'.

Floridi's GDI and its consequences give us what is currently the only formal and detailed analysis of the information concept of potential direct value for information science.

Finally, we will conclude this section by considering conceptions of knowledge of relevance for information science, several of which have already been mentioned by virtue of their link to the information concept.

Knowledge for information science

There are two main models – models, that is, in a very simple, conceptual and almost pictorial sense – used to describe the relation between information and knowledge in the information sciences. Though both are useful, they are incompatible. They differ in how the idea of knowledge is understood, and – since this is really a matter of how we use a label – it cannot be said that one could be right and the other wrong.

One model, stemming from Popper's epistemology, mentioned in the last chapter, uses 'knowledge' to denote World 2, the subjective knowledge within an individual person's mind. 'Information' is used to denote communicable knowledge, exchanged between people or recorded; this is Popper's World 3 of objective knowledge, necessarily encoded in a World 1 document, or physical communication. Information, in this model, is the form in which knowledge is communicated; 'knowledge in transit'.

The second model regards information and knowledge as the same kind of entity. Knowledge is seen as a form of 'refined' information, set into some form of larger structure, and both information and knowledge may be internal, in someone's mind, or external, encapsulated in some kind of document. This is usually presented as a linear progression, or a pyramid, from 'data', through 'information' to 'knowledge', perhaps with 'wisdom' or 'action' at the far end of the spectrum or the apex of the pyramid, as shown in Figure 4.3 on the next page; see, for example, Ackoff (1989), Rowley (2006; 2011) and Frické (2009). Some variations include 'capta' – data in which we are interested – between data and information (Checkland and Holwell, 1998). We have seen that both Bates and Floridi advocate views of information which allow for such a relation with data

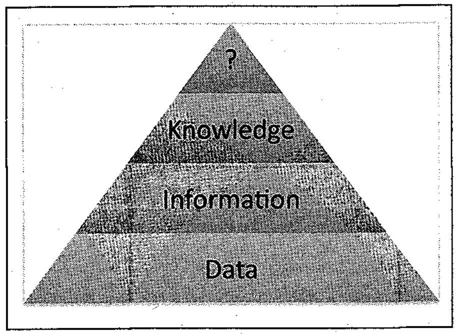


Figure 4.3 *The D-I-K-? hierarchy*

and with knowledge. Zins (2007) presents many definitions of data, information and knowledge, and their interrelation, given by participants in a Delphi study on the nature of information science.

If the incompatibility of these models causes concern, they can be integrated, but at the cost of having to distinguish two forms of knowledge: communicable, objective knowledge instantiated in documents, and personal, subjective knowledge in people's heads.

Knowledge has been studied by philosophers for many centuries, as the subject of epistemology. The usual view in that context is that it is to be understood as 'justified, true belief'; that is to say, for something to count as knowledge it must be believed by someone, for rational reasons, and it must be true.

This viewpoint poses problems for the information sciences (Buckland, 2012). When we think of knowledge, it is usually as the contents of collections of documents, rather than something which a particular person believes. The British Library, for example, has as its mission to 'advance the world's knowledge'; we doubt if they are concerned to discover who believes the contents of each item in their collections. oprávnenie Justification is also somewhat different in our content. Most philosophical argument justifies belief through sense perceptions; someone believes something because they saw something, heard something, etc. However, justification in a library and information context is likely to be based on information gained from some kind of document. This fits into epistemology in the form of svedectvo 'testimony'; a kind of evidence in which philosophers are becoming increasingly interested, and which overlaps with the ideas of social epistemology discussed in Chapter 3; see, for example, Audi (1997), Goldman (2009) and Adler (2010). Finally, there is a particular problem with the 'truth' criterion. Particularly with scientific and technical material, we can be sure that most of our current knowledge base is 'untrue', in the sense that it will be superseded in time, by better theories, more accurate observations, improved technologies, etc. It seems counter-intuitive to suggest, on this basis, that we have 'no knowledge'.

Some recent developments from philosophy have gone some way to provide new perspectives on knowledge, more appropriate for the information sciences. One has been mentioned already in previous chapters; this is the idea of Karl Popper of knowledge 'without a knowing subject', and consequently of a World 3 of objective knowledge which includes errors and inconsistencies. Popper's 'objective epistemology' (1979) views the totality of things – including information and its communication – as explicable by a system of three 'Worlds'. World 1 is the physical world, of people, books, computers, buildings, etc. World 2 is the internal, subjective mental state of an individual, including their personal knowledge. World 3 is the world of communicable, objective knowledge – or information – with which libraries and information centres deal. So, we may see the communication of information as involving all three of these 'worlds': a person may read a book (World 1), understanding the information it contains (World 3), and assimilating it into their own personal knowledge (World 2).

Other developments include:

- Luciano Floridi's (2010; 2011b) conception of knowledge in the Philosophy of Information, which, while generally respecting established philosophical ideas, adapts them to deal with the 'information context'.
- David Deutsch's (2011) concept of 'explanatory knowledge', which comprises our best rational explanations for the way the world is; such knowledge is inevitably fallible and imperfect, and our task is to improve it, not to justify it.
- Jonathan Kvanvig's (2003) idea of knowledge as 'understanding', allowing for contradictions and inconsistencies.
- Michael Polanyi's (1962) ideas of 'personal knowledge' (somewhat similar to Popper's World 2), which have acted as a basis for some conceptions of the practice of knowledge management; see, for example, Tsoukas (2005) and Day (2005).

It seems likely that some of these ideas will prove valuable in providing a precise understanding of knowledge in a form useful for our disciplines.

Finally, we should note that, although most consideration of knowledge for the information sciences has focused on the Western rational and scientific tradition, there has also been some interest in preserving the indigenous knowledge of peoples in the developing world and making it accessible, as we will discuss in Chapter 12. These forms of knowledge may differ radically from Western norms, being based in experiences, stories and even dreams, and typically having very different forms of classification of material and means of dissemination (Stevens, 2008; Maina, 2012).

We now turn to the next pair of fundamental concepts: documents and collections, dealing first with documents.

Documents

'The debate about whether "information" or "document" is the primary object of study in information science', writes Karamuftuoglu (2009), 'is a complex and multifarious one'. Throughout the debates about what information really is, there has been a strong trend which argues that this is not an important issue, since librarians and information scientists do not handle information at all, they handle documents. Certainly, these documents carry information, but they are not the same thing. And, as we have seen in earlier chapters, the documentation movement of the first half of the 20th century was one of the predecessors of information science.

The usual dictionary definition of 'document' is a paper, or other object, conveying some information, evidence or proof. The word comes from the Latin *docere*, to teach or to inform, with '-ment' implying the means.

It might therefore be thought that the meaning of 'document' is straightforward, and that the only issues that can arise relate to the differences between printed and electronic documents. This is far from the case. It can be argued that, if a 'document' is some physical thing which records thoughts or ideas – information – then we should include paintings, sculpture, and perhaps even any artefact, as documents. But then, what about geological specimens in museums, carefully labelled and collocated with similar minerals; might we not go to look at them to find out about the mineral, as an alternative to reading about it in a book – which is clearly a document. An antelope in the wild may be just an antelope; but in a zoo, placed where it can be studied, in surroundings indicative of its natural habitat, does it not convey information? And is an information-bearing 'thing' not a document? If a photograph of the lunar surface is a document, why not the real thing seen through a telescope? And so on.

The debates about what can count as a document have been very much associated with the work of Michael Buckland, whose views on 'information-as-thing' we noted earlier in this chapter. A British-born information scientist, who has spent much of his career as a distinguished professor in the United States, Buckland has espoused the centrality of the document idea for information science, looking backward to its origins in the documentation movement and forward to the newer ideas of document theory. A concise review of the whole topic and its history is given by Lund and Skare (2010).

Buckland brought the issues to attention in his influential 1997 article on 'What is a document?'. His views stem from his advocacy of 'information-as-thing', as against 'information-as-knowledge' or 'information-as-process' (Buckland, 1991), discussed earlier in this chapter, and hence the idea that a document is a physical entity. Reviewing and critiquing the views of 20th-century European documentalists, including Paul Otlet, Suzanne Briet and Donker Duyvis, he showed how they extended the idea of a document beyond printed texts, and indeed beyond items created with the intention of communicating information. The emphasis moved from communication to evidence: a document was some physical item which gave evidence, and was in that way informational. However, there had to be something else: to count as a document, the item had to be placed in relation to other evidence-bearing items, by being put in a collection, indexed, cross-referenced, etc. This is exemplified in the box here, using the examples of Briet quoted by Buckland.

Object	Document
star in the sky	No
photo of star	Yes
stone in river	No
stone in museum	Yes
animal in wild	No
animal in zoo	Yes

There are practical consequences of accepting this wider view of document: whole sub-disciplines such as museum documentation only make sense if it is accepted. The practice of botanical gardens, such as Kew Gardens in London, of arranging library books, dried herbarium specimens and living plants according to the same classification scheme is another example. The emphasis is on whether some item functions as a document, rather than what its physical form or nature may be.

Buckland's presentation has been reviewed and updated by several authors, including Buckland himself (1998), Frohmann (2009) and Kallinikos, Aaltonen and Marton (2010), particularly with respect to the specific issues of digital documents. These cannot be defined by form, medium or location in the way that physical documents can, being less 'solid' and permanent; hence the pragmatic 'documentation' approach, defining a document by function, is particularly appropriate. Analyses have been given of the document nature of religious icons (Walsh, 2012), museum objects (Latham, 2012), and even landforms (Grenersen, 2012). An intriguing new form of document is the *DNA barcode*, a small segment of genetic code, unique to a particular organism. They may be used to aid identification of all forms of living things, for example, identifying invasive organisms or illegal imports at national borders, or verifying validity of herbal remedies. They even hold out the prospect of a hand-held device, akin to *Star Trek*'s tricorder, which can process a sample to extract the sequence, and compare it with a remote database (Savolainen et al., 2005; Chase, 2008).

Even if we restrict ourselves to regarding documents as being some physical entity deliberately created for the purpose of conveying information, we still find some difficulties. In particular, we may have a problem deciding when two documents are 'the same'. If I have a copy of a textbook, and you have a copy of the same edition of the same book, then these are clearly different objects. But in a library catalogue they will be treated as two examples of the same document. What if the book is translated, word for word, as precisely as possible, into another language: is it the same document, or different? At what point do an author's ideas, on their way to becoming a published book, become a 'document'?

These are quite difficult issues philosophically, and pose real practical problems for resource description, in particular for library cataloguing codes. They led to the development of the Functional Requirements for Bibliographic Records (FRBR) model (Carlyle, 2006), discussed in the context of information organization in Chapter 6. FRBR distinguishes four levels:

- work: 'a distinct intellectual or artistic creation', e.g. Shakespeare's *Hamlet*
- expression: 'the intellectual or artistic realization of a work', e.g. the English text of *Hamlet*
- manifestation: 'the physical embodiment of an expression of a work', e.g. a specific edition of *Hamlet*
- item: 'a single exemplar of a manifestation', e.g. this copy of this edition of *Hamlet* in my hand.

However, there has been scope for disagreement on how these four levels should be understood (see, for example, Zhang and Salaba, 2009). Are adaptations or abridgements of an original to be considered as different expressions of the work, or as new works? When do small changes in a text, perhaps just the correction of errors when a book is reprinted, make a new expression, rather than just variations of the original expression? Do we need the idea of a 'superwork', to include commentaries, criticisms, different formats – e.g. a film based on a book? And so on. It seems that much of this ambiguity stems from the basic question of what is a document and, similarly, what is a 'work' (see, for example, Smiraglia, 2001).

Collections

Collections, in the sense of interest to the information sciences, and documents go naturally together. As we have seen, one of the things which makes a document a document is that is in an organized collection of related documents; and it is difficult to envisage a collection of information without documents.

> Collection: a number of objects collected or gathered together, viewed as a whole; a group of things collected and arranged.
>
> *Oxford English Dictionary*

The concept of a collection – an organized set of information-bearing items chosen for a particular purpose in a particular context or environment, and usually unique to that situation – is fundamental to the information sciences. Clayton and Gorman (2001, xii) describe a library collection as 'an assemblage of physical information sources combined with virtual access to selected and organized information sources'. The library is the archetypical form of information collection: as Corrall (2012, 3) puts it ' . . . the whole notion of a library is fundamentally associated with the idea of a collection, to the extent that the words "library" and "collection" are almost synonymous.' However, the same principles apply to collections in records centres, archives, museums, galleries, and all forms of information centre. For overviews, see Fieldhouse and Marshall (2012), Cullingford (2011) and Hughes (2012). Issues of collection management will be discussed in Chapter 12.

A useful distinction, though not an absolute one, is between two forms of collection:

- collections of ideas, embodied in documents
- collections of objects, which may provoke ideas in the viewer.

The former is typified by the library, the latter by the gallery.

It is the selection and organization of the collection, with items having some rationally assessed similarity, which distinguishes the typical information collection from the 'cabinet of curiosities' style of seemingly anarchic or random objects, intended to provoke surprise, awe or humour. (For thoughts on the relevance of the cabinet of curiosities to library and information science, see Divelko and Gottlieb, 2003 and Frohmann, 2009.)

With different forms of documents, and other items, and with other purposes, it forms the basis for the activities of information centres, libraries, records centres, and archives, museums (of all kinds, including natural history) and galleries (although the last might not consider the information-bearing nature of the items to be the most important). The disciplines which underlie these professional activities are often referred to as the 'collection sciences'.

The 'collection' in each of these cases may refer to the totality of their information resources, or to a sub-set: for example, a special collection within a library devoted to some topic, or the records of a particular organization within an archive centre.

The move towards a digital information environment has implications for the nature of the collection itself, as well as the practicalities of managing it.

Few collections are truly digital – examples of such would be the digital libraries created rapidly to deal with natural disasters – and these fully digital collections may be assembled very rapidly, in a way which physical collections cannot be. Conversely, very few collections now avoid any digital dimension. Managing 'hybrid' or 'complex' collections presents particular problems of maintaining consistency of approach, and presenting the collection in a unified way.

The move to digital collections generally means a move from 'ownership' and 'location' to 'access', as has been recognized for many years. The collection takes on a dynamic and impermanent nature, with varied and unpredictable lifecycles, and to some degree no longer under local control.

Any digital resource, perceived as a single entity by its users, may in fact be a composite or aggregate of distributed sources: a collection, composed of such entities is even more an aggregate. Given that the resources of which it is composed will often link to other resources, it becomes very difficult to place a 'boundary' around the collection, clearly distinguishing what is inside it. Furthermore, a single digital resource may be simultaneously in several digital collections, which could not happen in the pre-digital world.

This means that such a collection can no longer be thought of as a set of physical items in a particular location, or even a collection of digital items on a particular computer, but rather a means of selecting items from the universe of knowledge. The collection may be defined as 'a set of criteria for selecting resources from the broader information space', with collection membership defined 'through criteria rather than containment' (Lagoze and Fielding, 1998).

We will now consider the final set of fundamental concepts, those which arise largely from studies of information retrieval and seeking: relevance, aboutness and information use.

Relevance and aboutness

In an early paper, Wilson (1978) identified five 'fundamental concepts of information retrieval': information; aboutness; relevance; information need; and information use. More have been introduced over the years, to the point where Jansen and Rieh (2010) were able to present 17 'theoretical constructs of information searching and information retrieval'.

We have already considered the nature of information. The other concepts here fall into two groups: those concerned with the *aboutness* of a document, and its *relevance* in relation to a search or query; and those concerned with information needs and use. The concept of *relevance* has received particular attention; see for example Nolin (2009), Saracevic (1975; 2007), Borlund (2003) and Mizzaro (1997). As Jansen and Rieh (2010, 1525) say

> It is difficult to find a concept that has generated more discussion in or had more impact on the fields of information searching and information retrieval than has relevance. Relevance plays a most significant, fundamental and central role in all aspects of information retrieval and information searching, including theory, implementation and evaluation.

Tefko Saracevic, a Croatian-born professor of information science with a long career in the USA, is well known for his studies of relevance over three decades: he goes so far as to suggest that it is 'a, if not even *the*, key notion in information science in general and information retrieval in particular' (Saracevic, 2007, 915). This strong claim is based on the idea that much research and practice in information science is based on the simple idea that information systems provide answers to queries, and that each of these answers is relevant or not relevant to the question asked; in turn this is linked to the *aboutness* of the information retrieved. This concept, stating what a document is about – its subject matter – might appear straightforward, but has in fact been the topic of much debate; see Hjørland (2001) for an overview.

But it is the nature of relevance that has received most discussion over many years. 'Intuitively', says Saracevic (1975, 324) 'we understand quite well what relevance is. It is a primitive "y' know" concept . . . for which we hardly need a definition'. But when the idea is studied more closely for theory or research, it is not at all simple. There has even been controversy about such a basic issue as whether relevance is objective – a document is relevant to the query or it is not – or whether it is subjective – it is relevant if a particular user on a particular occasion thinks it is; see Hjørland (2004) and Abbott (2004).

However, it is in the evaluation of information retrieval systems that relevance becomes a complex and slippery concept, particularly when real users with real information needs are involved. Documents may be judged as relevant but not interesting, as irrelevant but useful, as relevant but not as relevant as other documents, as relevant if presented first in a results list but not relevant if presented last, and so on. The necessity to ignore such issues in laboratory-style investigations is one of the limitations of the systems paradigm, as discussed in Chapter 3.

To take an example used by one of the authors of this book many years ago (Bawden, 1990), suppose that the query is for information on 'treatment of migraine'. When the relevance of the results is considered, the first question is whether the documents are indeed about this topic; whether, in effect, the retrieval system including is indexing is working well. The second question is whether the user considers them relevant.

The first question is easily answered, by reading the documents. The second is more subtle. One document may be an account of recent advances in migraine treatment; the second may be review of the pharmacological action of a drug, or the biography of a famous doctor, with the treatment of migraine mentioned in passing. Although both are formally relevant – the system is working correctly in retrieving them – it is likely that most users would judge the first to be relevant and the second not; although if the aim were to compile a complete bibliography they might both be considered relevant. Further, users will often bring in other factors in deciding whether items are relevant, in the sense of being useful answers to their queries. The first document, although clearly the right topic, may be, for example, rather old, or written in a language which the user cannot read, or even written by an author with whom they have professional disagreements; and for any of these, or many other reasons, will be judged not relevant.

Furthermore, a user's opinion of the relevance of any particular document may change as the results are examined. The 'recent advances' document may be judged relevant if it is the first to be examined; but if several similar documents have already been seen, so that no new information is added, this decision may change. More generally, as a search progresses and it becomes clear what information is available on a topic, the user's view of what is relevant may change considerably. Taylor, Zhang and Amadio (2009) discuss this phenomenon, and give an example of the way in which relevance judgements changed as a group of students searched for information on a business topic. They found a variety of criteria other than simple aboutness being used, particularly the clarity of writing in the documents, and the students' ability to understand the information. For similar findings, see Taylor (2012).

The nature of relevance in particular contexts is also problematic: Raper (2007), for example, discusses geographic relevance, where the output from geographic information systems (GIS) must be judged in terms of place, space and time, as well as other attributes.

To deal with these complexities, a variety of frameworks for understanding different degrees and levels of relevance has been devised, and concepts such as pertinence, utility, usefulness and topicality have been advanced as alternatives to simple relevance; see, for example, Nolin (2009), Saracevic (2007), Borlund (2003), Mizzaro (1997), Bawden (1990) and Buckland (1983) for reviews and examples. These go a long way from the simple yes/no idea of relevance, originally devised for retrieval experiments: it has been suggested, for example, that relevance is best modelled by the mathematics of fractals (Ottaviani, 1994). There is still no completely satisfactory way of dealing with the relevance concept, either in theory or for practical system evaluation.

Information use and users

Turning to issues of information use and users, by contrast with the attention to the idea of relevance, the nature of *information use* has been very little studied, and there have been few attempts to give conceptual clarity to this very basic idea. In two of the few studies of this kind, Savolainen (2009) compares approaches to understanding information use, and Fleming-May (2011) analyses the related idea of *library use*. Other studies have looked at information use in terms of its context within the users' life and work, and at information practices and information for performing tasks; these will be discussed in the chapter on information behaviour.

These analyses have shown that the concept of information use is a complex and multifaceted one, and one which is changing in newer digital environments, particularly with the growth of social media. In particular, the rather restricted view of a 'user' as a passive recipient and consumer of certain limited types of information, typically held by information practitioners of the past, has been supplanted by the recognition that a much wider range of activities needs to be included; this view is typified by the suggestion that 'information users' would be better regarded as 'information players' (Nicholas and Dobrowolski, 2000).

Summary

All of the concepts discussed in this chapter are contested, and views of them have changed considerably over time.

Information and knowledge, and the relationship between them, are indeed slippery concepts. The significance of any link between the ideas of information in the physical, biological and social worlds remains unclear, while even within the information sciences there are alternative ways of regarding the concepts. Shannon's MTC has proved unsuitable as the basis for a formal treatment of information in our discipline; perhaps Floridi's GDI will be more successful. New approaches to knowledge may be more helpful for our purposes than traditional viewpoints.

Documents and collections are clearly entities of fundamental importance to the information sciences; but theory is weak here, and there are elements of ambiguity in how we understand these concepts, even before we consider the changes brought about by the transition to a largely digital information environment. Even ideas such as relevance and information user, whose meaning might be thought to be intuitively obvious, have complexities when analysed fully.

Further study of these fundamental entities and concepts can be of benefit to both research and practice.

- The fundamental concepts of the information sciences include information, knowledge, documents, collections, relevance and aboutness, and information use and users.
- All of these are contested concepts, whose nature and significance are still debated.
- Shannon's MTC has proved too limited as a theoretical basis for the information sciences; Floridi's GDI, and some of the newer concepts of knowledge, show promise in this respect.
- Further study of the fundamental concepts of the information sciences is important both for the academic subjects and the practical disciplines.

Key readings

Luciano Floridi, *Information – a very short introduction*, Oxford: Oxford University Press, 2010.

[A short and readable account of Floridi's approach, with concise reviews of other aspects.]

James Gleick, *The information: a history, a theory, a flood*, London: Fourth Estate, 2011.

[A readable account of many aspects of the information concept, centred around the development of Shannon's theory.]

Michael Buckland, What is a 'document'?, *Journal of the American Society for Information Science*, 1997, 48(9), 804–9.

[A classic review of the issue.]

Tefko Saracevic, Relevance: a review of the literature and a framework for thinking on the notion in information science. Part II: nature and manifestations of relevance, *Journal of the American Society for Information Science and Technology*, 2007, 58(13), 1915–33.

[A recent review of the issues.]

References

Abbott, R. (2004) Subjectivity as a concern for information science: a Popperian perspective, *Journal of Information Science*, 30(2), 95–106.

Ackoff, R. L. (1989), From data to wisdom, *Journal of Applied Systems Analysis*, 16(1), 3–9.

Adler, J. (2010) Epistemological problems of testimony, in Zalta, E. N. (ed.), *Stanford Encyclopedia of Philosophy* (Winter 2010 edition) [online] available at http://plato.stanford.edu/archives/win2010/entries/testimony-episprob.

Audi, R. (1997) The place of testimony in the fabric of knowledge and justification, *American Philosophical Quarterly*, 34(4), 405–22.

Barrow, J. D., Davies, P. C. W. and Harper, C. L. (2004) *Science and ultimate reality*, Cambridge: Cambridge University Press.

Bates, M. J. (2005) Information and knowledge: an evolutionary framework for information, *Information Research*, 10(4), paper 239 [online] available at http://informationr.net/ir/10-4/paper239.html.

Bates, M. J. (2006) Fundamental forms of information, *Journal of the American Society for Information Science and Technology*, 57(8), 1033–45.

Bawden, D. (1990) *User-oriented evaluation of information systems and services*, Aldershot: Gower.

Bawden, D. (2001) The shifting terminologies of information, *Aslib Proceedings*, 53(3) 93–8

Bawden, D. (2007a) Information as self-organised complexity: a unifying viewpoint, *Information Research*, 12(4), paper colis31, available at http://informationr.net/ir/12-4/colis/colis31.html.

Bawden, D. (2007b) Organised complexity, meaning and understanding: an approach to a unified view of information for information science, *Aslib Proceedings*, 59(4/5), 307–27.

Belkin, N. J. (1978) Information concepts for information science, *Journal of Documentation*, 34(1), 55–85.

Belkin, N. J. and Robertson, S. E. (1976) Information science and the phenomenon of information, *Journal of the American Society for Information Science*, 27(4), 197–204.

Borlund, P. (2003) The concept of relevance in IT, *Journal of the American Society for Information Science and Technology*, 54(10), 913–25.

Brillouin, L. (1956) *Science and information theory*, New York NY: Academic Press.

Buckland, M. K. (1983) Relatedness, relevance and responsiveness in retrieval systems, *Information Processing and Management*, 19(4), 237–41.

Buckland, M. K. (1991) Information as thing, *Journal of the American Society for Information Science*, 42(5), 351–60.

Buckland, M. K. (1997) What is a 'document'?, *Journal of the American Society for Information Science*, 48(9), 804–9.

Buckland, M. K. (1998) What is a 'digital document'?, *Document Numérique*, 2(2), 221-230, available from http://people.ischool.berkeley.edu/~buckland/digdoc.html [An amended version of Buckland's 1997 paper, with more emphasis on digital documents.]

Buckland, M. K. (2012) What kind of science can information science be?, *Journal of the American Society for Information Science and Technology*, 63(1), 1–7.

Capurro, R. and Hjørland, B. (2003) The concept of information, *Annual Review of Information Science and Technology*, 37, 343-411.

Carlyle, A. (2006) Understanding FRBR as a conceptual model: FRBR and the bibliographic universe, *Library Resources and Technical Services*, 50(4), 264–73.

Chase, M. W. (2008) Barcoding life, in *McGraw Hill Yearbook of Science and Technology 2008*, New York NY: McGraw Hill, 35–7.

Checkland, P. and Holwell, S. (1998) *Information, systems and information systems – making sense of the field*, Chichester: Wiley.

Clayton, P. and Gorman, G. E. (2001) *Managing information resources in libraries: collection management in theory and practice*, London: Library Association Publishing.

Cole, C. (1994) Operationalizing the notion of information as a subjective construct, *Journal of the American Society for Information Science*, 45(7), 465–76.

Cole, C. (1997) Calculating the information content of an information process for a domain expert using Shannon's mathematical theory of communication: a preliminary analysis, *Information Processing and Management*, 33(6), 715–26.

Cornelius, I. (2002) Theorising information for information science, *Annual Review of Information Science and Technology*, 36, 393–425.

Corrall, S. (2012) The concept of collection development in the digital world, in Fieldhouse, M. and Marshall, A. (eds), *Collection development in the digital age*, London: Facet Publishing, 3–25.

Cullingford, A. (2011) *The special collection handbook*, London: Facet Publishing.

Davies, P. and Gregersen, N. H. (2010) *Information and the nature of reality: from physics to metaphysics*, Cambridge: Cambridge University Press.

Day, R. E. (2005) Clearing up 'implicit knowledge': implications for knowledge management, information science, psychology, and social epistemology, *Journal of the American Society for Information Science and Technology*, 56(6), 630–5.

Deutsch, D. (2011) *The beginning of infinity: explanations that transform the world*, London: Allen Lane.

Divelko, J. and Gottlieb, K. (2003) Resurrecting a neglected idea: the reintroduction of library museum hybrids, *Library Quarterly*, 73(2), 160–198.

Fieldhouse, M. and Marshall, A. (eds) (2012) *Collection development in the digital age*, London: Facet Publishing.

Fleming-May, R. A. (2011) What is library use? Facets of concept and a typology of its application in the literature of library and information science, *Library Quarterly*, 81(3), 297–320.

Floridi, L. (2010) *Information – a very short introduction*, Oxford: Oxford University Press.

Floridi, L. (2011a) Semantic conceptions of information, in Zalta, E.N. (ed.), *Stanford Encyclopedia of Philosophy (Spring 2011 edition)* [online] available at http://plato.stanford.edu/archives/spr2011/entries/information-semantic.

Floridi, L. (2011b) *The philosophy of information*, Oxford: Oxford University Press.

Frické, M. (2009) The knowledge pyramid: a critique of the DIKW hierarchy, *Journal of Information Science*, 35(2), 131–42.

Frohmann, B. (2009) Revisiting 'what is a document?', *Journal of Documentation*, 65(2), 291–303.

Furner, J. (2010) Philosophy and information studies, *Annual Review of Information Science and Technology*, 44, 161–200.

Genersen, G. (2012) What is a document institution? A case study from the South Sámi community, *Journal of Documentation*, 68(1), 127–33.

Gleick, J. (2011) *The information: a history, a theory, a flood*, London: Fourth Estate.

Goldman, A. I. (2009) Social epistemology: theory and applications, in O'Hear, A. (ed.), *Epistemology: Royal Institute of Philosophy Supplement 64*, Cambridge: Cambridge University Press, 1–18.

Hjørland, B. (2001) Towards a theory of aboutness, subject, topicality, theme, domain, field content . . . and relevance, *Journal of the American Society for Information Science and Technology*, 52(9), 774–8.

Hjørland, B. (2004) Arguments for philosophical realism in library and information science, *Library Trends*, 52(3), 488–506.

Hjørland, B. (2007) Information: objective or subjective/situational? *Journal of the American Society for Information Science and Technology*, 58(10), 1448–56.

Hughes, L. M. (ed.) (2012) *Evaluating and measuring the value, use and impact of digital collections*, London: Facet Publishing.

Jansen, B. J. and Rieh, S. Y. (2010) The seventeen theoretical constructs of information searching and information retrieval, *Journal of the American Society for Information Science and Technology*, 61(8), 1517–34 .

Kallinikos, J., Aaltonen, A. and Marton, A. (2010) A theory of digital objects, *First Monday*, 15(6), available from http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/3033/2564.

Karamuftuoglu, M. (2009) Situating logic and information in information science, *Journal of the American Society for Information Science*, 60(1), 2019–31.

Kvanvig, J. L. (2003) *The value of knowledge and the pursuit of understanding*, Cambridge: Cambridge University Press.

Lagoze, C. and Fielding, D. (1998) Defining collections in distributed digital libraries, *D-Lib Magazine*, (November 1998) [online] available from http://www.dlib.org/dlib/november98/lagoze/11lagoze.html.

Latham, K. F. (2012) Museum object as document: using Buckland's information concepts to understand museum experiences, *Journal of Documentation*, 68(1), 45–71.

Liebenau, J. and Backhouse, J. (1990) *Understanding information*, London: Macmillan.

Lund, N. W. and Skare, R. (2010) Document theory, *Encyclopedia of Library and Information Sciences* (3rd edn), 1:1, London: Taylor & Francis, 1632–9.

Ma, L. (2012) Meanings of information: the assumptions and research consequences of three foundational LIS theories, *Journal of the American Society for Information Science and Technology*, 63(4), 716–23.

Machlup, F. and Mansfield, U. (1983) *The study of information; interdisciplinary messages*, New York NY: Wiley.

Maina, C. K. (2012) Traditional knowledge management and preservation: intersections with library and information science, *International Information and Library Review*, 44(1), 13–27.

Mizzaro, S. (1997) Relevance: the whole history, *Journal of the American Society for Information Science*, 48(9), 810–32.

Nicholas, D. and Dobrowolski, T. (2000) Re-branding and re-discovering the digital information user, *Libri*, 50(3), 157–162.

Nolin, J. (2009) 'Relevance' as a boundary concept: reconsidering early information retrieval, *Journal of Documentation*, 65(5), 745–67.

Ottaviani, J. S. (1994) The fractal nature of relevance: a hypothesis, *Journal of the American Society for Information Science*, 45(4), 263–72.

Polanyi, M. (1962) *Personal knowledge*, Chicago IL: University of Chicago Press.

Popper, K. R. (1979) *Objective knowledge: an evolutionary approach (revised edition)*, Oxford: Clarendon Press

Raper, J. (2007) Geographic relevance, *Journal of Documentation*, 63(6), 836–52.

Rowley, J. (2006) Where is the wisdom that we have lost in knowledge?, *Journal of Documentation*, 62(2), 251–70 .

Rowley, J. (2011) The wisdom hierarchy: representations of the DIKW hierarchy, *Journal of Information Science*, 33(2), 163–80.

Saracevic, T. (1975) Relevance: a review of and a framework for the thinking on the notion in information science, *Journal of the American Society for Information Science*, 26(6), 321–43.

Saracevic, T. (2007) Relevance: a review of the literature and a framework for thinking on the notion in information science. Part II: nature and manifestations of relevance, *Journal of the American Society for Information Science and Technology*, 58(13), 1915–33.

Savolainen, R. (2009) Information use and information processing: comparison of conceptualizations, *Journal of Documentation*, 65(2), 187–207.

Savolainen, V., Cowyn, R. S., Vogler, A.P., Roderick, G. K. and Lane, R. (2005) Towards writing the encyclopaedia of life: an introduction to DNA barcoding, *Philosophical Transactions of the Royal Society: Biology*, 360(1462), 1805–11.

Shannon, C. E. and Weaver, W. (1949) *The mathematical theory of communication*, Urbana IL: University of Illinois Press.

Smiraglia, R. P. (2001) *The nature of a 'work': implications for the organization of knowledge*, Lanham MD: Scarecrow Press.

Stevens, A. (2008) A different way of knowing: tools and strategies for managing indigenous knowledge, *Libri*, 58(1), 25–33.

Stonier, T. (1990) *Information and the internal structure of the universe*, Berlin: Springer-Verlag.

Stonier, T. (1992) *Beyond information: the natural history of intelligence*, Berlin: Springer-Verlag.

Stonier, T. (1997) *Information and meaning: an evolutionary perspective*, Berlin: Springer-Verlag.

Taylor, A. (2012) User relevance choice and the information search process, *Information Processing and Management*, 48(1), 136–53.

Taylor, A., Zhang, X. and Amadio, W. J. (2009) Examination of relevance criteria choices and the information search process, *Journal of Documentation*, 65(5), 719–44.

Tsoukas, H. (2005) *Complex Knowledge: studies in organizational epistemology*, Oxford: Oxford University Press.

Vedral, V. (2010) *Decoding reality: the universe as quantum information*, Oxford; Oxford University Press.

Von Baeyer, C. (2004) *Information: the new language of science*, Harvard MA: Harvard University Press.

Walsh, J. A. (2012) 'Images of God and friends of God': the holy icon as document, *Journal of the American Society for Information Science and Technology*, 63(1), 185–94.

Wilson, P. (1978) Some fundamental concepts of information retrieval, *Drexel Library Quarterly*, 14(2), 10–24.

Zhang, Y. and Salaba, A. (2009) *Implementing FRBR in libraries: key issues and future directions*, New York NY: Neal-Schuman.

Zins, C. (2007) Conceptual approaches for defining data, information and knowledge, *Journal of the American Society for Information Science and Technology*, 58(4), 479–93.

CHAPTER 5

Domain analysis

> If information science is to be taken seriously as a field of study, it is important that basic theories are formulated and examined in the field. Domain analysis is one serious attempt to consider the basic problems in IS. Anybody working in the field should care about the arguments that have been or might be raised for or against this view.
>
> Birger Hjørland (2010, 1653–4)

Introduction

Domain analysis is a 'metatheoretical framework for library and information science . . . the basic claim in [domain analysis] is that "domains" of knowledge are the proper object of study for LIS' (Hjørland, 2010, 1648). It is also a very practical framework for understanding information on particular topics, and for particular groups, and underlies the work of the subject specialist information practitioner. It provides a valuable link between research and practice in the information sciences.

The idea was conceived by, and remains closely associated with, Birger Hjørland, Professor of Information Science at the Royal School of Librarianship and Information Science, Copenhagen. Hjørland – whose work on the philosophy and concepts of information science we have already encountered in earlier chapters – has for many years been one of the foremost authorities on the foundations of information science and of knowledge organization.

We will use Hjørland's meaning of domain analysis here; but note that the phrase has sometimes been used in the library and information literature with a more restrictive meaning, usually relating either to bibliometric analysis or to classification.

Domain analysis as a theory for information science

As we saw in Chapter 3, the idea of domain analysis was initially introduced in association with the socio-cognitive paradigm, for information science, as an alternative to the cognitive and behavioural paradigms (Hjørland and Albrechtsen, 1995). These latter approaches focused on the individual, and their personal knowledge, behaviours, preferences, information needs, opinions of relevance, etc. By contrast, wrote Hjørland and Albrechtsen,

> The domain analytic paradigm in information science (IS) states that the best way to understand information in IS is to study the knowledge domains as thought or discourse communities . . . Knowledge organization, structure, co-operation patterns, language and communication forms, information systems, and relevance criteria are reflections of the objects of the work of these communities and of their role in society.

They acknowledge that this is not an entirely new idea, and that previous approaches have shared its assumptions: as Tennis (2003) puts it 'domain analysis is done in many ways and by many people in information science'. It had not, however, previously been stated fully or explicitly. Talja (2005) gives early examples of this kind of approach.

As understood by these authors, domain analysis is a realist approach in philosophical terms; it seeks a basis for information science in factors external to the individual, which are objective rather than subjective, and which may be located in the expertise and practices of subject specialists. If we wish to design an information system for Scandinavian geography, Hjørland and Albrechtsen say, the obvious approach is to design it according to how Scandinavia actually is, not according to the way some particular users think it is; we would probably want to use geographers as the relevant domain experts to advise us. However, it is not necessary to adhere to this philosophical viewpoint to make pragmatic use of domain analysis; see, for example, Feinberg (2007) for a more subjective approach.

They also argue that domain analysis, being based on groupings of people with common interests and concerns, is primarily a social theory, and that this implies that information science is primarily a social science. (As we noted in Chapter 1, we interpret this as meaning a field of study, with a focus on information in a social context.) More specifically, this is a *socio-cognitive* approach, since it considers the communication of knowledge within groups of people:

> Domain analysis consequently does not conceive users in general, but sees them as belonging to different cultures, to different social structures, and to different domains of knowledge. Information producers, intermediaries, and users are more or less connected in communities that share common languages, genres and other typified communication practices.
>
> Hjørland (2010, 1652); for a fuller discussion, see Hjørland (2002a)

This theoretical background to domain analysis has influenced views of the nature of information science, and its practice: as Sundin (2003) puts it 'domain analysis has, during the last decade, developed as an important theoretical approach within library and information science'. And, in particular, as we saw in Chapter 1, we can understand information science as the application of the methods of domain analysis to the information communication chain (Robinson, 2009).

Domain analysis also has considerable practical value for the information scientist, and for information specialists generally, as we will see later.

What is a domain?

In straightforward pragmatic terms, an information domain is the set of information systems, resources, services and processes associated with a group of users with common concerns and a common viewpoint, and sharing a common terminology. This will typically be: an academic subject area, e.g. theoretical physics or philosophy; a professional or trade, e.g. accountancy or clock-making; or an 'everyday' hobby or concern, e.g. cookery or job-seeking.

There are some overlaps and complications. Medicine, law and engineering, for example, are all professions and also academic subjects. History may be the concern of the professional historian, the student of history (at any level, from junior school to university), and the layperson (either as a recreational interest, or for some specific purpose, e.g. researching family history). Domains may be defined generally or specifically, with this in mind.

More formally, Hjørland (2010, 1650) suggests that a domain may be 'a scientific discipline or a scholarly field. It may also be a discourse community connected to a political party, a religion, a trade or a hobby'. Domains, he suggests, are defined and explained by three dimensions: ontological, epistemological and sociological. The ontological dimension defines the domain by its main object of interest: botany by plants, history by the past, theology by the Divine, etc. This is the most usual way of defining a domain. The epistemological dimension relates to the kind of knowledge in the domain, or perhaps different kinds of knowledge associated with different paradigms or ways of understanding. The sociological dimension relates to the kind of people and groups involved in the domain.

These three dimensions interact in a rather complex way, and it is necessary to produce something more concrete in order to apply domain analysis to a specific case. Hjørland (2002b, 2010) suggests that 11 'aspects' or 'approaches' to the study of domains can be derived from the three dimensions, as discussed in the next section.

Other commentators have extended these ideas. Tennis (2003) argues that use of Hjørland's 11 approaches does not delineate exactly what a domain is, in any particular case. He adds two 'axes', to help definition of domains:

1 'Areas of modulation' set parameters on the names and extension of the domain, specifying what is included and not included, and what the domain

may sensibly be called (e.g. is Transpersonal Psychology different from psychology *per se*, or is it a part of mainstream psychology?).

2 'Degrees of specialization' set the 'intension' of a domain, i.e. the focus of its specialization (Hinduism, for example, is a qualified domain of religion, with lesser extension (scope) and greater intension (specialized focus), and also indicates intersection, when domains overlap (e.g. medical ethics).

See Hjørland and Hartel (2003), and Feinberg (2007), for further discussion of how a domain is described, or constructed.

Sundin (2003) extends the domain analytical approach by using tools from the theory of professions, looking at professional interests, power relations, and occupational identities, in an examination of the professional domain of nursing. Morado Nascimento and Marteleto (2008), in similar vein, take ideas from Bourdieu's sociology of culture to better understand the nature of domains, and the rationale for information practices within particular domains.

One might also add to the specified 11 aspects: including, for example, issues of the evaluation and analysis of information, and standards and quality and reliability issues, which differ considerably between domains.

These debates imply that domain analysis must be treated as a general approach, rather than a precise algorithm: as Hartel (2003) puts it, investigations are conducted 'in the general domain analytical spirit'.

Aspects of domain analysis

Hjørland (2002b) introduced 11 'aspects' of domain analysis; essentially things which feature in study and practice of information with specific domains. They are expressed in slightly different words in various publications, but are summarized in the box here.

Aspects of domain analysis

1 Resource guides and subject gateways
2 Special classifications and thesauri
3 Indexing and retrieval special features
4 User studies
5 Bibliometric studies
6 Historical studies
7 Document and genre studies
8 Epistemological and critical studies
9 Terminologies, languages for special purposes, discourse analysis
10 Structures and institutions in communication of information
11 Cognition, knowledge representation and artificial intelligence.

It is worth noting that these approaches straddle the boundary between what is generally regarded as 'research' (e.g. user studies or bibliometrics) and what is usually regarded as 'professional practice' (e.g. producing literature guides, and knowledge organization tools). This is an attractive feature of the approach, as it emphasizes the desirable links between research and practice in the information sciences.

It is worth taking each of these aspects in turn, and examining its nature with some examples. It should be noted that this is not a definitive list, and other aspects could be added, though it has been satisfactory for most analyses carried out in practice.

Resource guides

The preparation and use of resource guides stems from the long-standing involvement of information specialists in subject bibliography and guides to the literature, albeit that it is now most commonly expressed in the form of resource guides and subject gateways in digital formats. Such guides may be generally applicable and publicly available, or may be for in-house use, relating to sources available within a particular institution. There is a strong relation between this aspect of domain analysis and the promotion of digital literacy for specific subjects, discussed in Chapter 13. An examination of existing literature guides is an essential part of preparation for information work in the area. Analysis of the context and structure of such guides, particularly their development over time, is a valuable contribution to understanding the nature of the information domain.

Information organization tools

Similarly, the creation and use of classifications, taxonomies, thesauri and other tools for information organization, which will be discussed in Chapter 6, has been the concern of information specialists for many decades. 'Creation' is most likely to be concerned with small, specialized and possibly in-house tools – taxonomies and thesauri for the most part. However, some information scientists will become involved with the updating and maintenance of subject-specific sections of major tools, such as the Dewey, UDC, Bliss and Library of Congress classification schemes, and the Library of Congress subject headings. As with literature guides, a familiarity with relevant tools is essential for any subject specialist information practitioner, while analysis of the nature and development of these tools sheds much light on information in particular domains.

Indexing and retrieval

The same is true of specialized indexing and retrieval systems, whose development and use is the concern of information specialists in those subject areas. Specialist retrieval systems, discussed further in Chapter 7, are most common in science, technology and medicine (STEM) subject areas, but are also found elsewhere. Examples are:

- chemical information systems, allowing for retrieval of substances, reactions and properties
- molecular biology systems, allowing matching of nucleotide and protein sequences
- medicines information systems, which allow identification of medicines by their actions, their appearance, and their names, the latter allowing of language variants and phonetic matching
- geographic information systems, with spatial and temporal metadata and displays
- fine arts, with image databases allowing retrieval by subject, genre and copyright and use status.

Sophisticated and specialist indexing is also often associated with STEM subjects, though is found elsewhere. Examples are:

- medicine and pharmaceuticals, with an array of indexing languages
- fine arts, which have developed exhaustive indexing languages, which also serve as informative glossaries
- archaeology and history, for which many 'local' vocabularies have been developed.

User studies

These have been a feature of the library and information landscape for several decades and are discussed in Chapter 9. Many have focused on professional or academic groups ('information needs of doctors', 'information behaviour of lawyers', 'information use by students of business and finance', etc.), and hence are relevant to the idea of information use within specific domains (Case, 2012). However, as Hjørland (2002b) points out, their value in domain analysis is limited in many cases by the lack of any theoretical basis, or investigation of any domain specific aspects; there is limited value in asking the same questions of senior business managers as of new students of geology. Studies of users and use within a particular domain may be very valuable, if the study focuses on the specific resources, tasks and knowledge of relevance to that domain.

Within a subject area, of course, there may be several very different user groups, and hence potentially different domains. Users of mathematics literature, for example, a subject with a particularly great 'reach', include: professional mathematicians; professionals in closely associated areas, for example statistics and operations research; those working in the many areas which use mathematics, hence the plethora of 'mathematics for . . . ' books; those teaching and learning mathematics, at any level from junior school to graduate school; and those interested in recreational mathematics. While these may most sensibly be counted as different domains, they are all in some sense 'users of mathematics information'.

It is of obvious importance for information scientists working in a subject area to be familiar with what is known of the information practices, behaviours and needs of those involved with that area. It may be feasible for them to carry out small and local user studies, ideally building on previous similar studies, to avoid results only of local interest.

Bibliometrics

Bibliometric studies of the literature of a subject area are similar to user studies, in that information practitioners will be more likely to make use of their results, rather than carry out such studies for themselves; although local small-scale bibliometric studies may be valuable if they can be combined or compared with other results. Bibliometric data will be valuable to any subject-specialist practitioner, in scoping the size and nature of the information base of a subject, and in assisting in collection development, as discussed in Chapter 8.

Historical perspectives

Historical studies within domain analysis comprise two distinct, though related, aspects: the study of the historical development of the subject area itself, and of its concepts, theories and practices; and the historical development of its information resources, systems and services. It is not possible to understand the second without an appreciation of the first. This forms a part of the more general study of information history, discussed in Chapter 2.

Historical aspects, as we saw in that earlier chapter, are sometimes regarded as of little relevance to practice. For domain analysis, we can make the same case for history as was made generally: that we cannot properly understand current information provision, and plan for its improvement, without understanding how it has come to be as it is.

Document and genre studies

Studies of documents and genres are of evident relevance to information practitioners in specific domains, since each domain has a typifying mix of types of document and content which defines it, very often uniquely. The changes in the types of document available as the information environment becomes largely digital, as noted in Chapter 4, makes it even more important for these factors to be studied and understood.

Some subjects have forms of documents and content very closely associated with them, e.g. chemistry (chemical structures and reactions), history (archival and primary sources), music (sheet music and recorded sound), fine arts (images and artefacts), geography (maps and atlases), astronomy (star charts), mathematics (mathematical notations), architecture (plans, drawings and models), etc. Others are typified by a very wide range of documents and content; healthcare for example, has a particularly large and diverse range of resources (Robinson, 2010).

Epistemological and critical studies

These focus on the nature and structure of knowledge in specialist subject areas. It is evident that the kind of knowledge being expressed, and the way it is communicated, differs greatly between, say, mathematics, the visual arts, and horticulture. This affects the kind of information resources which are provided, and the way in which information and knowledge are accessed. For example, Robinson (2010) notes the importance of three kinds of knowledge in the healthcare domain: *propositional knowledge*, publicly available, objective, and often of a scientific or technical nature; *practical craft knowledge*, tacit and gained by professional experience, and often associated with 'clinical intuition' and 'professional judgement'; and *personal knowledge*, subjective and gained by reflection on experience. For a fuller discussion, see Higgs, Richardson and Dahlgren (2004).

As Hjørland (2002b) emphasizes, some subject areas may have several paradigms, or schools of thought. It is important that information providers are aware of these, as otherwise information provision may be partial or confusing.

Terminology, language, discourse

The importance of terminology and special languages, and discourse analysis, will differ greatly between areas, although almost all domains, even 'everyday' or hobbyist topics, have some specific terms, or words used with a particular meaning. The way in which these terms are used, and more generally the way discourse is carried on, is also characteristic of subject areas, and needs to be appreciated by those involved with information provision.

The issue is most evident in the terminology of STEM subjects in particular, with vocabularies such as chemical nomenclature, giving unambiguous names for chemical substances, botanical and zoological nomenclatures for naming living things, and the terminologies of the medical sciences, being the most obvious. These special languages form a barrier to any information access from the outside, and one concern of the information practitioner may be to provide access in layperson's language, particularly for healthcare (Robinson, 2010).

These areas also often generate artificial languages, largely or wholly divorced from natural language: mathematical notation, chemical structure notation, and (in a different kind of domain altogether), musical and dance notations, are examples. We might also include meta-languages, such as the Unified Medical Language System (UMLS), used for purely formal information processing and vocabulary linking rather than person-to-person communication here.

However, the issue is by no means confined to these subjects: a scan of any academic bookstore will show a variety of dictionaries and glossaries for subjects such as philosophy, law, history, economics and the fine arts.

Information practitioners in subject domains have an obvious need for familiarity with any specific terminologies used, and may be able to contribute by creating and updating glossaries and similar tools. They may also be involved with another aspect of language; the translation of terminology between natural languages, for subject-specific language dictionaries or multilingual glossaries.

Structures, institutions and organizations

An understanding of how these are involved in the communication of information within a particular domain is of evident importance to information practitioners. This involves participants at all stages of the communication chain from producers, through all forms of disseminators, to users: research institutes and universities; learned societies, commercial and institutional publishers, libraries, archives, etc. These factors may differ from domain to domain in perhaps unexpected ways: to give just one example, the role of learned societies in journal publishing and database production is much more significant in mathematics than in similar subjects.

Cognition, knowledge representation and artificial intelligence (AI)

This final aspect is perhaps the least likely to have direct relevance for the information practitioner. This is particularly so since the 1980s enthusiasm for 'expert systems' able to encapsulate the knowledge in a specific domain and hence of great interest to subject specialist information scientists, has produced very little of practical value. Nonetheless, research continues on many aspects of knowledge representation, and it is important for subject specialists to be aware of developments in their areas. Perhaps the most relevant current aspects are ontologies and meta-languages such as UMLS mentioned above.

We can therefore see that of the 11 aspects, some – for example, resource guides and information organization tools – are largely regarded as practically useful activities rather than research topics, while for others – for example, knowledge representation, bibliometrics and user studies – the opposite is the case. In fact, however, all have a two-fold value, as being both tools for the practitioner and also topics for study and research. In this way, domain analysis forms a bridge between theory and practice for information science.

Practical value of domain analysis

Hjørland explains the aspects of domain analysis as being the special competencies of the information scientist; they encompass the skills and knowledge needed to provide effective information systems and services in particular subject areas, and for particular groups of users.

They therefore form the basis for subject specialist information work, as will be discussed fully later in this chapter.

One may also, in a sense, reverse this idea, and note that aspects of domain analysis may be of value in studying a subject area in information terms. If we wish to know what an academic subject, professional discipline, etc., is like 'informationally', we will want to know, for example: what kind of information and knowledge does the area encompass?; what kind and format of information resources are available, and who produces them?; what kind of special terminologies, classifications, indexes, and retrieval systems are available?; what guides to the literature, subject gateways, etc., are available?; how much is published on the topic, in what languages, and so on?; how old is the subject, and how important is older material? – and so on. All the questions are covered by the aspects of information domains.

Domain analysis will therefore be of practical importance in three ways, and to three different groups. Experienced practitioners in an area will apply its aspects: creating and using resource guides, creating and using knowledge organization tools, etc. New practitioners, or those new to a subject area, can use the aspects to gain competence: finding out what special retrieval tools are available, assessing what is known of users and their needs, etc. And researchers can use the aspects of domain analysis as a framework for examining a domain in information terms; since if one knows about all the aspects as they relate to an area, one can realistically claim to understand it. Domain analysis therefore provides a clear and direct link between research and practice in information science.

It should be noted that even the most enthusiastic proponents of domain analysis do not believe that information science research and practice should split into domain-specific subjects: there are general information science principles applicable to all domains (Hjørland, 2010).

We will now look at some specific examples of the use of domain analysis to understand information in a specific subject area.

Examples of domain analysis

The most extensive published example of domain analysis is that of Robinson (2010), who analyses the healthcare domain – which has a particularly rich and diverse set of resources and users – using all of the 11 aspects. The analysis is structured into six main sections: domain overview, including epistemological and structure or institution aspects; history; producers and users; information organization, including classifications, terminologies and knowledge representation; sources and retrieval systems, including resource guides, analysis of document types and bibliometric analysis; and information and knowledge management. This is an indication of the way in which the 11-aspect framework can be used as guide, rather than a prescriptive formalism.

Other studies which have applied some of the aspects of domain analysis to specific subjects include Hartel (2003; 2010) on cookery, Karamuftuoglu (2006) on information art, Morado Nascimento and Marteleto (2008) on architecture, Orom (2003) on fine art, and Sundin (2003) on nursing. Talja (2005) reviews earlier examples of this general approach, before the specific domain analysis formulation was described.

Domain analysis and the subject specialist

Information practitioners have taken subject specialist roles for many years. Examples include: subject specialist librarians in university or special libraries; subject specialist cataloguers in national or research libraries; information officers in biomedical research institutes or pharmaceutical companies; information researchers in business and financial institutions; and so on.

A subject specialist information practitioner is not a subject specialist *per se*: a medical information specialist is not a doctor, nor a legal information specialist a lawyer (unless they have qualified in those professions previously; it is not uncommon, particularly in commerce and industry, for such positions to be filled by subject practitioners who have switched to an information role). As Hjørland (2010, 1649) puts it: 'To be an information specialist with a given speciality is not to be a subject specialist in the ordinary sense, but rather to be an expert in information resources in that field'.

What is needed is someone with an understanding of, at least, the basic concepts of the subject – its 'logic and language' – plus specialist and deep knowledge of its information attributes, which are well expressed in the domain analysis approaches. There has always been a debate as to whether it is better to first gain subject knowledge, through degree level study and perhaps practical experience, and then to add on an understanding of, and skills in, the information aspects; or to begin with a grounding in information science, and add on the subject knowledge. But the two must both be acquired at some stage. It is essential to combine some subject knowledge, plus knowledge of subject-specific sources, users needs, etc.: the sort of aspects noted above. (Rodwell (2001) refers to this as 'subject expertise'; Hjørland (2000) as the capabilities of the 'domain generalist'. The subject background is needed to make possible the detailed insight into needs and sources, to enable the interpretation and evaluation of information, and to make the information professional credible to their users.

What does such a subject specialist do that other information professionals cannot (or prefer not to)? A variety of answers are given (see, for example, Rodwell, 2001; Hardy and Corrall, 2007; Garitano and Carlson, 2009; Jackson, 2010), most commonly:

- create user guides
- create terminologies and taxonomies
- carry out 'difficult' searches and reference queries
- evaluate and interpret information
- provide current awareness
- suggest useful new sources and collection development
- act as instructor, trainer, consultant and advisor.

The resonance with the aspects of domain analysis, discussed above, is clear.

This often leads to an overlap with the activities of the users, and with the organization as a whole: in an academic library, for example, this takes the form of an involvement with teaching. The downside, if the subject specialist role is over-emphasized, can be a concern about loss of professional identity and skills.

There has been some diminution in the perceived importance of subject specialism for the information practitioner since 1990, largely due to the view that the most important task for most information practitioners was to provide access to digital resources. The limitations of this view are now being realized; Hjørland (2010) gives a powerful argument for the need for subject knowledge by most, if not all, information practitioners. This will best be achieved by bringing domain analysis to the forefront as a main focus for research and practice in information science.

Summary

We have seen that the idea of domain analysis is a central concept for information science, linking several other important aspects: resources and retrieval systems, terminologies and classifications, user behaviour, the quantitative aspects of a literature as assessed by bibliometrics, the nature of knowledge in a subject area, the way information sources and services have developed over time, the institutions and organizations that produce and disseminate information, etc.

By focusing on particular subject areas and user groups, it provides both a theoretical framework and a set of specific activities and competencies, for researchers and practitioners in the information sciences. This gives a unique approach to both the study of, and the practical instantiation of, the information communication chain; and thereby provides a unique stance for information science, distinct from adjacent disciplines such as computer science and information systems and publishing.

- Domain analysis is a framework for studying information communication within subject areas and user groups, and for the provision of information services to such groups.
- It is a socio-cognitive approach, based on examining the nature of knowledge within social groups, and implications for information provision.
- It provides a basis for the work of the subject specialist information practitioner.
- It provides a bridge between research and practice in the information sciences.

Key readings

Birger Hjørland, Domain analysis in information science, in *Encyclopedia of Library and Information Science* (3rd edn), 1:1, 1648–54, Abingdon: Taylor & Francis, 2010. [A concise overview of the ideas and applications.]

Birger Hjørland, Domain analysis in information science. Eleven approaches – traditional as well as innovative, *Journal of Documentation*, 2002, 58(4), 422–64. [Explanations and examples of the eleven original aspects of domain analysis.]

References

Case, D. O. (2012) *Looking for information: a survey of research on information seeking* (3rd edn), Bingley: Emerald.

Feinberg, M. (2007) Hidden bias to responsible bias: an approach to information systems based on Haraway's situated knowledge, *Information Research*, 12(4), paper colis07, available from http://InformationR.net/ir/12-4/colis/colis07.html.

Garitano, J. R. and Carlson, J. R. (2009) A subject librarian's guide to collaborating on e-science projects, *Issues in Science and Technology Librarianship*, no. 57, available from http://www.istl.org.

Hardy, G. and Corrall, S. (2007) Revisiting the subject librarian: a study of English, law and chemistry, *Journal of Librarianship and Information Science*, 39(2), 79–91.

Hartel, J. (2003) The serious leisure frontier in library and information science: hobby domains, *Knowledge Organisation*, 30(3/4), 228–38.

Hartel, J. (2010) Managing documents at home for serious leisure: a case study of the hobby of gourmet cooking, *Journal of Documentation*, 66(6), 847–76.

Higgs, J., Richardson, B. and Dahlgren, M. A. (eds) (2004) *Developing practice knowledge for health professionals*, London: Butterworth-Heinemann.

Hjørland, B. (2000) Library and information science: practice, theory and philosophical basis, *Information Processing and Management*, 36(3), 504–31.

Hjørland, B. (2002a) Epistemology and the socio-cognitive perspective in information science, *Journal of the American Society for Information Science and Technology*, 53(4), 257–70.

Hjørland, B. (2002b) Domain analysis in information science. Eleven approaches – traditional as well as innovative, *Journal of Documentation*, 58(4), 422–64.

Hjørland, B. (2010) Domain analysis in information science, in *Encyclopedia of Library and Information Science* (3rd edn), 1:1, 1648–54, Abingdon: Taylor and Francis.

Hjørland, B. and Albrechtsen, H. (1995) Toward a new horizon in information science: domain-analysis, *Journal of the American Society for Information Science*, 46(6), 400–25.

Hjørland, B. and Hartel, J. (2003) Afterword: ontological, epistemological and sociological dimensions of domains, *Knowledge Organisation*, 25(4), 162–201.

Jackson, M. (2010) *Subject specialists in the 21st century library*. Oxford: Chandos.

Karamuftuoglu, M. (2006) Information arts and information science: time to unite?, *Journal of the American Society for Information Science and Technology*, 57(13), 1780–93.

Morado Nascimento, D. and Marteleto, R. M. (2008) Social field, domains of knowledge and informational practice, *Journal of Documentation*, 64(3), 397–412.

Orom, A. (2003) Knowledge organization in the domain of art studies – history, transition and conceptual changes, *Knowledge Organization*, 30(3/4), 128–43.

Robinson, L. (2009) Information science: communication chain and domain analysis, *Journal of Documentation*, 65(4), 578–91.

Robinson, L. (2010) *Understanding healthcare information*, London: Facet Publishing.

Rodwell, J. (2001) Dinosaur or dynamo? The future for the subject specialist reference librarian, *New Library World*, 101(1), 48–52.

Sundin, O. (2003) Towards an understanding of symbolic aspects of professional information: an analysis of the nursing domain, *Knowledge Organisation*, 30(3/4), 170–81.

Talja, S. (2005) The domain analytic approach to scholars' information practices, in Fisher, K. E., Erdelez, S. and Mckechnie, L. (eds), *Theories of information behavior*, Medford NJ: Information Today, 123–7.

Tennis J. T. (2003) Two axes of domains for domain analysis, *Knowledge Organisation*, 30(3/4), 191–5.

CHAPTER 6

Information organization

> The first step in wisdom is to know the things themselves: this notion consists in having a true idea of the objects: objects are distinguished and known by classifying them methodically and giving them appropriate names. Therefore, classification and name-giving will be the foundation of our science.
>
> Carl Linnaeus

> Cataloguers would lose much of their status if it were shown that most cataloguing is a trivial job easily done by clerical staff.
>
> Maurice Line

Introduction

The organization of information, and information resources, is one of the fundamental aspects of the information sciences. In essence, this amounts to classifying and name-giving: as essential in our sciences as Linnaeus proclaimed it to be in his, though for rather different reasons. It is an extensive and complex subject in its own right, but fortunately it has a particularly wide range of textbooks and articles. We will outline the main topics and issues within the subject, pointing to where more detailed treatments can be found. One of the main aspects of the subject is the way in which relatively old tools and techniques are being adapted to the modern information environment. Therefore, although up-to-date materials are important, older texts may also be very useful. The fundamentals of the subject, particular some of the theory of classification and indexing, go back many years, and – as we saw in Chapter 2 – some of the main tools used today have their origins in the 19th century. They were developed as part of the attempt to provide bibliographic control of printed materials; to record, identify and make accessible all the intellectual output of humanity, as expressed in recorded knowledge. They, and newer equivalents, are now being used to provide access to the rapidly expanding stock of digital material.

We will look first at some of the fundamental issues of information organization, before examining the main tools: terminology, metadata, resource description, systematic and alphabetic subject description and abstracting. Texts covering several of these aspects in detail include Chowdhury and Chowdhury (2007), Taylor (2004), Chan (2007) and Svenonious (2000). For a more detailed examination of all aspects of information organization in one subject domain – healthcare – see Robinson (2010, Chapter 4).

This subject is referred to as either 'information organization' or 'knowledge organization'. Usually these terms are treated as synonymous. But they remind us that the purpose may be either understanding the structure of knowledge itself, or the pragmatic purpose of arranging documents; initially physical documents on shelves, latterly digital documents in a virtual space. The questions raised have been asked since the earliest days of philosophy: is there a single 'natural' classification of everything in the world?; on what basis do we assign things to categories?; how do our mental concepts relate to physical things?; how distinct can two things be, and still be given the same name?; and so on. There is a great deal of theory underlying all organizations of information and knowledge; whether classification schemes for documents or scientific taxonomies – of plants, animals, rocks, stars, and many more.

For an overview of the theoretical bases for documentary information organization, see Svenonious (2000), Tennis (2008) and Hjørland (2003, 2008a), and for theoretical underpinnings of various aspects, see Hjørland (2008b, 2009, 2011), Hjørland and Pedersen (2005), Hjørland and Shaw (2010), and Weinberg (2009).

Controlled vocabulary and facet analysis

These are two fundamental concepts, both of which appear in several aspects of information organization.

A controlled vocabulary is, at its simplest, just a list of terms which are to be used for indexing and retrieval; examples are keyword lists, subject headings, classifications, taxonomies, thesauri, authority lists, and others. They 'control' the variability and redundancy of natural language. The opposite is an uncontrolled vocabulary: full-text, freely chosen keywords and tags, etc. The argument about whether controlled or uncontrolled vocabularies are 'best' has rumbled on for many years, quite pointlessly. The obvious answer is that a combination of both is desirable: controlled vocabulary for consistency, and to use term relations; uncontrolled for precision, and for new terms.

Facet analysis involves dividing the concepts within a subject domain into consistent sections. For example, the subject of 'historic buildings' might have the facets PURPOSE (house, church, school . . .), STYLE (Gothic, classical, Arts and Crafts . . .), CONSTRUCTION (stone, brick, timber . . .), AGE (Victorian, medieval . . .), etc. This style of analysis finds use in the construction of classifications and thesauri, the design of interfaces, the construction of complex search logics, and more (La Barre, 2010; Broughton, 2006a).

We can now begin to look at the tools for information organization, beginning with the most familiar: terminology lists.

Terminologies

A *terminology* or *term list* is generally taken to mean the words and phrases used in the communication of information in a particular subject or field; interpreted broadly, it includes words and phrases in common usage. Terminology resources encompass a number of overlapping categories: general dictionaries, scholarly, introductory or abridged, and illustrated; multilingual dictionaries; dictionaries and glossaries for specific subjects; special-purpose dictionaries, of rhyming words, quotations, crossword clues; and thesauri, in the sense of Roget's word finder, rather than the retrieval thesauri discussed below. We might also include dictionaries of the proper names of persons ('biographical dictionaries') or places (geographical dictionaries or gazetteers). Once the epitome of the quality-assured printed volume, these kinds of tools are now increasingly digital in format, and beginning to be produced by crowd-sourcing, particularly for popular use.

There is some overlap between subject terminology tools, such as glossaries, and the subject heading lists, thesauri and taxonomies discussed below. Both kinds may be used to provide terms for indexing, and equally both may be used to help understand the 'logic and language' of a subject. There is also overlap between the 'people and places' lists and the authority files used in cataloguing and resource description. They are primarily intended to fulfil the requirements noted by Buckland (2008) for the 'general reference' function: to provide, or confirm, facts (in this case, to explain the meaning or significance of a word or phrase), and to provide context for such explanation. For the origins and development of sources of this kind, see Hitchens (2005), Hüllen (2004), Mersky (2004) and Mugglestone (2005); and for current issues, see Tackabery (2005), Mugglestone (2011), Cassell and Hiremath (2011, Chapters 7, 10 and 11), and Ayre, Smith and Cleeve (2006).

Beyond subject specific dictionaries and glossaries there are also, particularly in STEM subjects, a variety of special languages and notations. Some notations, such as those of mathematics, music and dance, are languages in their own right. Others are detailed and specific terminologies, such as: nomenclatures for chemical structures and reactions (Leigh, 2011); nomenclatures and taxonomies for living things (Bowker, 2005; Heidorn, 2011), and specialized terminologies for medicines and healthcare (Robinson, 2010, Chapter 4).

The varied terminologies noted above can all provide terms for indexing, but their main purpose is communication, rather than information retrieval. Now we turn to vocabularies and standards specifically designed for indexing and retrieval, beginning with metadata, which provides 'standard containers' for descriptions of documents.

Metadata

Literally 'data about data', metadata is best understood as short, structured and standardized descriptions of information resources. The term first gained wide use in the 1990s, but the idea had been instantiated in library cataloguing rules in the mid-19th century. For an accessible introduction to metadata principles see Haynes (2004); for more detail of specific formats and schemas see Zeng and Qin (2008), Miller (2011) and Hider (2012).

Metadata records are surrogates for the original items, used in its place. The main purposes for metadata are as means to identify, retrieve, use and manage information resources. This includes: retrieval, finding required items by searching or browsing; display, deciding whether an item is likely to be useful; legal issues, noting the rights status of items; and records management, noting who has responsibility for the document, and when they should be reviewed, archived, etc. It is also desirable that metadata can be shared and exchanged; hence the requirement for standardization.

Metadata has usually been thought of as having two components: *descriptive metadata* and *subject metadata*. These were traditionally instantiated in different physical forms: for example, the distinct 'author/title catalogue' and 'subject catalogue', in libraries using drawers of catalogue cards, but both are now usually subsumed within one metadata format. Descriptive metadata describes the item itself – its title, author, date of publication or creation, physical form, etc. Subject metadata describes the content – what the item is 'about'. Subjects may be described using controlled terms – classification codes, subject headings and so on – or uncontrolled terminology – for example, terms from titles and abstracts. In terms of Popper's '3 Worlds', we can think of descriptive metadata as defining World 1 information objects, while subject metadata defines their World 3 content. Descriptive metadata answers questions such as 'what is this called?', 'who wrote it?', 'how old is it?', 'how big is it?', 'what kind of thing is it?', 'where is it?' Subject metadata answers the question 'what is it about?' The details of descriptive metadata will vary according to the physical nature of the item: e.g. 'how big is it?' will be answered in terms of pages and centimetres for a book, and megabytes for a digital resource. Subject metadata is invariant to physical form, since it describes the intrinsic content, and the same tools are therefore used regardless of form; e.g. the Dewey Decimal Classification, best known for shelf arrangement of books in libraries, is also used in internet directories.

As noted above, metadata records should be relatively short, and must be structured and standardized. By structured is meant, in effect, that a metadata record comprises fields, elements or attributes representing a distinct type of information, e.g. a title, a keyword, or an author name. In this way, we can distinguish records referring things written by Benjamin Disraeli from things written about him.

By standardized is meant that the same information is presented in the same way. Disraeli's name always appears as 'DISRAELI, BENJAMIN', or as 'BENJAMIN DISRAELI', or as 'DISRAELI B'. None is right or wrong, but there should be consistency.

As an aside, we might remember that Disraeli became Lord Beaconsfield in later life, and wrote some books under that name. It is good if our metadata collection has a link between the two. But to make such a link, two kinds of knowledge are needed: the general knowledge that it is possible for one person to have two names; and the specific knowledge that Disraeli/Beaconsfield is such a person. Human experts are good at amassing and using these kinds of knowledge, computers less so. Although automatic metadata creation is highly desirable, particularly to cope with the great quantities of digital information now available, intellectual metadata creation by expert people is still regarded as the best option to get the best quality; even though, as the opening quote from Maurice Line reminds us, even some within the information professions have doubted the mystique which has arisen about some aspects.

Some examples of currently important metadata standards which govern the content and elements of records are briefly noted below, to give an idea of their variety; for more details, see Zeng and Qin (2008) and Miller (2011).

- Machine Readable Cataloguing (MARC) is an exchange format for metadata records created according to the AACR2 and RDA cataloguing codes. In the past, there have been numerous national variants, e.g. USMARC and UKMARC, and international versions, particularly UNIMARC. MARC21 is now emerging as a new de facto international standard. Numbered fields and sub-fields are used to hold the record content.
- Dublin Core (DC) and derivatives – the best known web metadata format, DC was designed as the simplest possible useful metadata format, initially comprising just 15 elements – such as 'title', 'creator', 'subject' and 'format' – with minimal instructions for entering content. It has been expanded since, and numerous variants have been produced. One such is the UK government's e-GMS standard, which expands DC to 23 fields, including some appropriate for government material, such as 'mandate' and 'preservation'.
- Learning Object Metadata (LOM), designed to hold metadata for educational resources at all levels of granularity, from a diagram to a video clip to a textbook, and having a complex set of fields, including educationally specific elements such as 'typical age range' and 'interactivity type'.
- Text Encoding Initiative (TEI), a standard for the representation of texts in

digital form, particularly valuable for digital humanities. A series of tutorials and examples for TEI, 'TEI by example', is available at the time of writing at http://tbe.kantl.be/TBE.
- Metadata Encoding and Transmission Standard (METS), developed by the Library of Congress for archiving digital items; a modular and flexible standard. Metadata Object Description Schema (MODS) is an extension to METS to represent library cataloguing records.
- International Standard for Archival Description (General) (ISAD(G)), a metadata standard providing general guidance for description of archival records.
- Visual Resources Association (VRA) Core, a standard for description of works of visual cultures and images which document them.

These kinds of metadata standards and formats, which permit the creation of records describing information resources, are supported by a variety of languages and standards which allow their encoding and implementation in web environments; this is often seen as a move towards the *semantic web*, to be discussed in the next chapter. An important example is XML, in which several of the standards above, including LOM, METS, MODS and DC variants, are encoded. They are typically expressed as an XML *schema* with a *namespace* denoting the location of components and definitions.

The Resource Description Framework (RDF) is a general metadata model, based on the description of resources as a series of subject-relation-object expressions termed *triples*, very similar to the entity-relation database models discussed in the next chapter. It is regarded as the standard knowledge representation language for the Web. Topic maps are a rather similar kind of formal knowledge representation.

RDF and XML are used to build more specific metadata models, such as the Web Ontology Language (OWL), designed to represent ontologies, and the Simple Knowledge Organization System (SKOS), a model for representing controlled vocabularies such as classification schemes, taxonomies, subject heading lists and thesauri.

For an overview, see Antoniou and van Harmelen (2008), and as examples, see XML used to encode MARC records (Dimic, Milsavljevic and Surla, 2010), and an ontology represented in RDF to create a dataset of the sales of artworks (Allinson, 2012).

We will now look at how the first form of metadata content is provided: the description of an information item.

Resource description and cataloguing

This is the provision of descriptive metadata; presenting the physical form of a document. The term *cataloguing* still generally refers to library material; *resource description* is more general, and increasingly used. For detailed background on cataloguing, see Bowman (2003) and Welsh and Batley (2012).

Charles Ammi Cutter, the American librarian who was one of the 19th-century originators of modern ideas of resource description, argued that a catalogue should fulfil the following functions (phrased in more modern idiom):

- to allow a user to *find* a resource, for which they know one or more of author, title or subject
- to *show* what resources are available written by particular authors or on specified subjects
- to *help* the user choose the best resource for their needs, by *edition* (date, publisher, etc.) and by *character* (style, level, etc.).

These are still very relevant aims. A commonly quoted set of aims for catalogues of printed libraries was:

- *location* – identifying where particular resources are to be found
- *collocation* – bringing associated works (e.g. by the same author, or on the same subject, or in a series of books or reports) together
- *information* – providing directly some needed information (e.g. a full bibliographic reference, the full name of an author, the exact name of a corporate author).

These are also still relevant, even in a digital environment, if we think of resources being brought together dynamically on screen, rather than physically on a shelf.

These sets of general aims have been expanded into lists of more specific 'principles', which can guide the creation of explicit cataloguing codes, and other protocols for resource description. They also serve the desire for universal bibliographic control, by proving a consistent description for all published documents. See Bowman (2006) for an account of early developments. Most modern library cataloguing codes stem from the influential 'Paris Principles', approved by an international conference on cataloguing principles in 1961. The latest set of such principles is IFLA's 'Statement of International Cataloguing Principles' (ICP), which is inclusive of more types of resource than earlier principles (Tillett and Cristán, 2009); for a commentary on this, see Guerrini (2009).

Cataloguing codes, used in libraries for many decades, include very detailed rules for the precise description of such things as authors' names and editions of a work, and for specification of such things as the names of illustrators and translators, influenced by the need for precise identification of specific printed documents, particularly books, in large collections. The best known of these codes is AACR (Anglo-American Cataloguing Rules). The printed form of the 1998 revision of the second edition of these rules (AAČR2) includes 26 chapters over 676 pages: an indication of the necessary complexity of such rules.

AACR is based upon a more general standard, the ISBD (International Standard Bibliographic Description), dating from the early 1970s, which dictates at a more general level what can be said in describing a bibliographic item. This requires the following elements to be specified: title and statement of responsibility ['author']; edition; material, or type of publication ['physical form']; publication, distribution, etc. ['publisher']; physical description [size, etc.]; series; note; standard number and terms of availability. AACR specifies, in considerable detail for consistency, how these are to be expressed.

The limitations of cataloguing codes such as AACR led to the development of the Functional Requirements for Bibliographic Records (FRBR) model, which is beginning to make an impact on cataloguing systems and practices (IFLA, 1998; Zhang and Salaba, 2009). It is based on an entity-relation model, of the kind discussed in the databases section of the next chapter, which shows the relations between different kinds of documents, and their attributes, and the people and organizations which create and disseminate them; an example is shown in Figure 6.1.

As was mentioned in Chapter 4, FRBR distinguishes four levels – work, expression, manifestation, and item – but these are not entirely satisfactorily defined. Nonetheless, there is some evidence that this is a 'natural' way of looking at bibliographic entities (Pisanski and Žumer, 2010), as well as resting on a formal model.

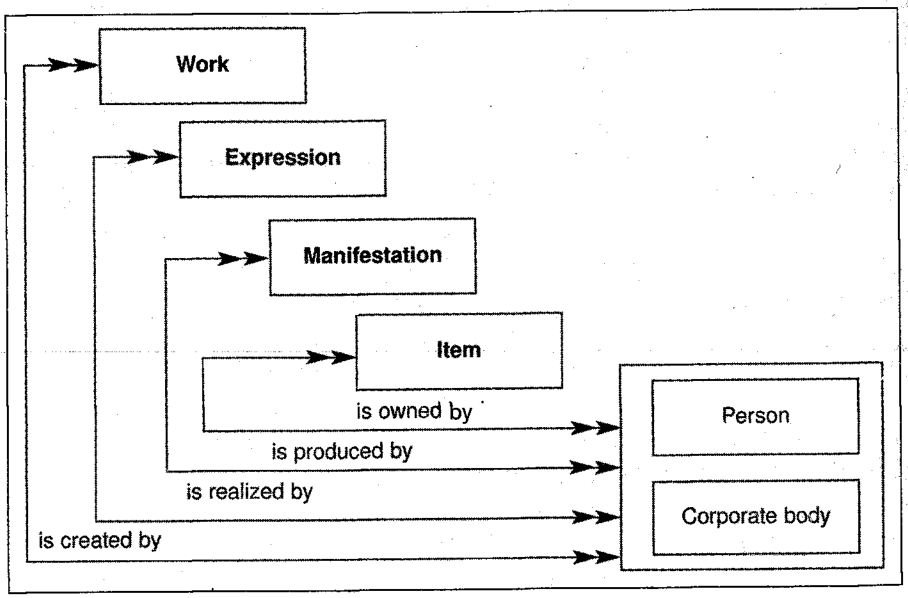


Figure 6.1 *Part of the FRBR model (reproduced from Chowdhury, 2010)*

Typically, library cataloguing has allowed searching at the 'manifestation' level, and display of results at the 'item' level. The FRBR model should allow for more flexible search and display, to allow for greater precision – e.g. 'retrieve records for this edition of this book, where there is a copy available for loan in a library to which I have access' – and also with more generality – e.g. everything on Shakespeare's *The Tempest*: the play itself, in its various editions, novels based on it, translations, commentaries, audiobook versions, and film versions, including related films, such as *Forbidden Planet* and *Prospero's Books*.

However, there has been disagreement on how these four levels should be understood, stemming from the basic questions of what is a document (discussed in earlier chapters) and what is a 'work' (Smiraglia, 2001).

FRBR, and the associated Functional Requirements for Authority Data (FRAD), which provides standard forms for the names or people, organizations, works, etc. (Patton, 2009), and Functional Requirements for Subject Authority Data (FRSAD) for subject description (Salaba, Zeng and Žumer, 2011), form the 'conceptual foundation' for the cataloguing standard which is intended to replace AACR2: Resource Description and Access (RDA). It is intended to be used beyond the library world, in the wider collection environment; in museums and archives, for example. For overviews, see Oliver (2010) and Anhalt and Stewart (2012).

Unlike AACR, RDA has been designed around a formal data model, and it avoids the complexity of separate rules for describing different kinds of material. However, it has been designed for compatibility with AACR, so that existing catalogue records need not be modified. It has been criticized from its inception as too much influenced by AACR, and hence stuck in the past; see, for example, Coyle and Hillman (2007). Its implementation was delayed, and it is still uncertain to what extent major libraries and information services will support it in its current form (Anhalt and Stewart, 2012).

We will now look at the tools used for describing subjects; the 'aboutness' of a document. Following the usual convention, we will divide them into *systematic* and *alphabetic* tools; classifications and taxonomies, and subject headings and thesauri, respectively. However, we should note that there is some overlap between the two; classification schemes often have alphabetic indexes, so that the place/s of specific subjects in the scheme can be quickly found, while alphabetic vocabularies which show broader and narrower terms can be drawn out as a taxonomy. We should also note the importance of tools which can map or translate between different vocabularies, linking the way a concept is treated in each. These are sometimes termed *crosswalks*, if they convert between two specific vocabularies or metadata format. There are also *metavocabularies*, which link several vocabularies or terminologies; an example is the Unified Medical Language System (UMLS), which links healthcare vocabularies (Robinson, 2010).

But first, we need to briefly look at a rather imprecisely used term; ontology.

Ontologies

There is no single, generally accepted meaning of 'ontology' within the information sciences. The term, coming from the Greek *ontos*, 'that which exists', has been used, and still is used within philosophy to mean the study of what kinds of things can exist, and how they can be described. It is sometimes used to describe, informally, a general set of things of concern; the ontology of an area in this sense would be the major concepts within it.

The term has been adopted within computer science to mean a formal description of a domain of knowledge, in terms of the entities within it, and their relationships. Understood in this way, many controlled vocabularies – classification schemes, taxonomies and thesauri in particular – would be regarded as ontologies, albeit rather simple ones, and they are often described as such, particularly in the context of the semantic web. The term is sometimes reserved for vocabularies with a rich set of relationships, more than the synomym/hierarchy/associative relations common to retrieval vocabularies. For example, biomedical ontologies use relations such as 'is contained in', 'adjacent to', 'preceded by', 'transformation of', and 'has participant' (Smith et al., 2005).

Systematic vocabularies: classification and taxonomy

Classifications, categorizations and taxonomies are all forms of knowledge organizations which aim to show the relationships between concepts – generally hierarchical relationships – by bringing together terms representing similar meanings. They are therefore referred to as 'systematic' vocabularies: constructed according to a 'system'. For detailed overviews of classification, both theory and practice, see Broughton (2005), Hunter (2009) and Bowker and Star (2000); a very clear brief account of principles, including applications in a web environment, is given by Slavic (2011).

The theory of classification can get very complex (see, for example, Langridge, 1992; Beghtol, 2010; Bowker, 2005; Bowker and Star, 2000; Hjørland and Pedersen, 2005; and Hjørland, 2008b), but some pragmatic points can be stated simply:

- Classification shows the relations between concepts; particularly, though not exclusively, hierarchical concepts.
- Classification is a process of categorizing concepts into mutually exclusive sets, by rational principles of division.
- Particular items can rarely be classified absolutely, but rather on the basis of overall similarity.

To be sensible and usable, a classification must:

- apply to similar things
- give sets similar in nature and size
- apply consistent criteria for division
- apply one criterion at a time.

These principles are always obeyed in formally designed classifications, but may be broken down into simpler taxonomies and categorizations.

All classifications have some form of *notation*; a numeric or alphanumeric code which reflects the structure of the classification, and shows the relationship between its component parts. For example, in the Dewey Decimal Classification the number 425 is used for 'grammar of standard English'. The decimal notation structure shows that this is part of class 420 'English and Old English', itself part of class 400, 'Language'. Classificatory notation has the advantage of being language-independent; a book on a particular topic will have the same notation in any library in the world which uses that classification.

There is some interrelation between systematic and alphabetic vocabularies. An alphabetic indexing vocabulary which shows broader and narrower terms in detail can be displayed as a hierarchical classification, while the 'top terms' of a thesaurus or set of subject headings can be used as a set of broad categories or a taxonomy.

Classification is a similar process to indexing; however, it is usual for just one classification notation to be assigned. This is not strictly necessary, unless the intention of classification is to provide a single place for the physical location of an item. For computerized material, as many classification codes as appropriate can be applied.

The simplest form of systematic vocabulary is a *categorization*, a simple form of classification, with limited structure and detail. 'Broad' categorizations are so called because they include wide subject concepts within each category. These are useful for browsing and for physical arrangement.

They are used in several settings: in libraries, particularly public or school libraries where a simple structure with a high degree of browsability is required, or where fiction is a large component of stock; in bookshops (both in the high street and online); and as an additional search tool in some computerized databases, particularly in the humanities and social sciences. They are also present in many web directories and portals.

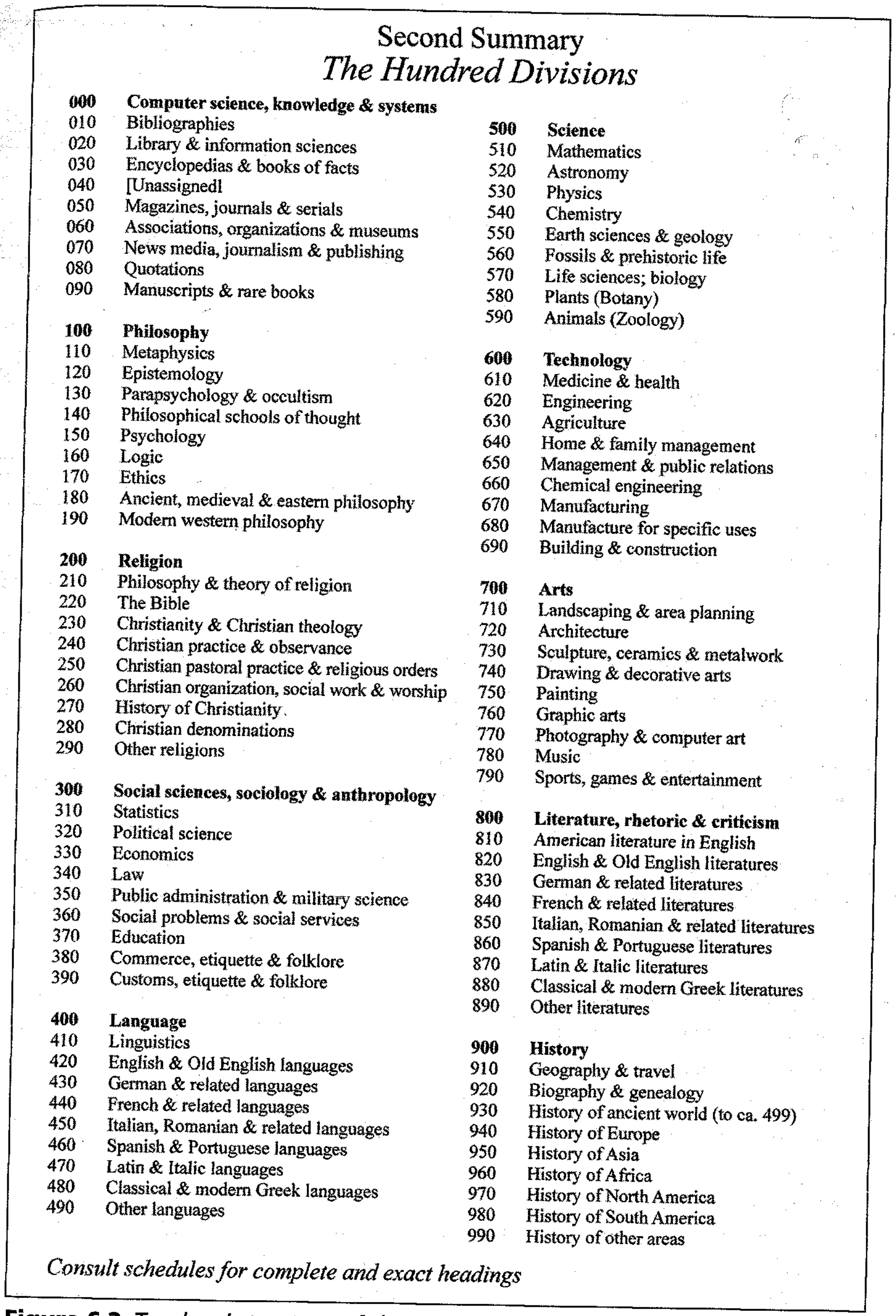
These are a simple form of classification, with little or no hierarchical structure, rarely going down more than one level, and often breaking the rules about single principle of division. For example, a bookshop may well have a general category for COOKERY, sub-divided into such things as VEGETARIAN COOKERY, CHINESE COOKERY, MICROWAVE COOKERY, etc. It is unlikely that it would divide to further levels, CHINESE VEGETARIAN COOKERY, etc., and it is unlikely that the shop owner would be worried by the inconsistent division by ingredients, region, cooking instruments etc. Such systems work because they are generally small-scale – browsers in a bookshop can see much of the stock and categories at one glance – and are geared closely to the needs of the users and the nature of the material. The same argument justifies their use for simple navigating access to a collection of digital material.

More complex is a *taxonomy*: a classification devised for a particular environment or set of information. They usually reflect the local conditions closely, and are adapted to the local 'culture'. They are usually modified and extended more frequently than other information organization tools. See Lambe (2007) for an overview of this kind of taxonomy. The term is, of course, also still used in an older sense: a scientific classification of some aspect of the natural world. Taxonomies are usually intellectually constructed, but may be automatically generated. They are a popular tool for organizing digital information resources, always supporting browsing, and sometimes searching as well.

Taxonomies may well 'break the rules' of classification, for example by allowing the same concept to appear at different levels, if this is helpful to the user's browsing. They may be a high-level description of subject matter – for example a 'corporate taxonomy' is a way of expressing the interests of an organization, so as to, for example, organize material on an intranet. This kind of taxonomy may link into a thesaurus for more detailed terminology. Alternatively, taxonomies may include many specific examples – names of places, people or departments, for example, as well as general concept headings – and may have more similarity with a thesaurus than a library classification. They may also include much descriptive information about the concepts and items, and be a kind of information-giving tool. Taxonomies are typically used to give access to diverse forms of materials – databases, documents, e-mails, people – and hence to support knowledge management programmes.

More complex and all-encompassing are the oldest established vocabularies of this kind: *enumerative* classifications, devised for physical arrangement of library materials, and more recently used also for subject retrieval of digital information. Enumerative classifications aim to list completely – to enumerate – all aspects of knowledge within their scope; they are invariably hierarchically arranged, dividing and subdividing knowledge; the Dewey classification, for example, is divided into ten main classes, each class into ten divisions, and each division into ten sections, with further sub-division as needed; its structure at the divisions level (with headings simplified for clarity) is shown in Figure 6.2, typifying this kind of enumerative classification.

They are well suited for arranging large volumes of material, especially when a physical arrangement, with a place for each item, is required, and hence are



Second Summary
The Hundred Divisions

000	**Computer science, knowledge & systems**
010	Bibliographies
020	Library & information sciences
030	Encyclopedias & books of facts
040	[Unassigned]
050	Magazines, journals & serials
060	Associations, organizations & museums
070	News media, journalism & publishing
080	Quotations
090	Manuscripts & rare books
100	**Philosophy**
110	Metaphysics
120	Epistemology
130	Parapsychology & occultism
140	Philosophical schools of thought
150	Psychology
160	Logic
170	Ethics
180	Ancient, medieval & eastern philosophy
190	Modern western philosophy
200	**Religion**
210	Philosophy & theory of religion
220	The Bible
230	Christianity & Christian theology
240	Christian practice & observance
250	Christian pastoral practice & religious orders
260	Christian organization, social work & worship
270	History of Christianity
280	Christian denominations
290	Other religions
300	**Social sciences, sociology & anthropology**
310	Statistics
320	Political science
330	Economics
340	Law
350	Public administration & military science
360	Social problems & social services
370	Education
380	Commerce, etiquette & folklore
390	Customs, etiquette & folklore
400	**Language**
410	Linguistics
420	English & Old English languages
430	German & related languages
440	French & related languages
450	Italian, Romanian & related languages
460	Spanish & Portuguese languages
470	Latin & Italic languages
480	Classical & modern Greek languages
490	Other languages
500	**Science**
510	Mathematics
520	Astronomy
530	Physics
540	Chemistry
550	Earth sciences & geology
560	Fossils & prehistoric life
570	Life sciences; biology
580	Plants (Botany)
590	Animals (Zoology)
600	**Technology**
610	Medicine & health
620	Engineering
630	Agriculture
640	Home & family management
650	Management & public relations
660	Chemical engineering
670	Manufacturing
680	Manufacture for specific uses
690	Building & construction
700	**Arts**
710	Landscaping & area planning
720	Architecture
730	Sculpture, ceramics & metalwork
740	Drawing & decorative arts
750	Painting
760	Graphic arts
770	Photography & computer art
780	Music
790	Sports, games & entertainment
800	**Literature, rhetoric & criticism**
810	American literature in English
820	English & Old English literatures
830	German & related literatures
840	French & related literatures
850	Italian, Romanian & related literatures
860	Spanish & Portuguese literatures
870	Latin & Italic literatures
880	Classical & modern Greek literatures
890	Other literatures
900	**History**
910	Geography & travel
920	Biography & genealogy
930	History of ancient world (to ca. 499)
940	History of Europe
950	History of Asia
960	History of Africa
970	History of North America
980	History of South America
990	History of other areas

Consult schedules for complete and exact headings

Figure 6.2 *Top-level structure of the Dewey classification (reproduced from Bowman, 2005)*

widely used for library classification. They have limitations in dealing with very detailed subject description, and with items involving several concepts; the extent to which they cover all of knowledge has also been critiqued (Zins and Santos, 2011). They are also unsuitable for rapidly changing subject fields, since they cannot be revised frequently. Topics such as 'Internet' and 'AIDS/HIV', which rapidly generated large volumes of literature, caused problems for enumerative classifications, which initially had no place for them. Because they are widely used internationally, their revision is generally in the hands of international committees, which produce revisions of particular sections and subsections at infrequent intervals.

Nonetheless, enumerative classifications are still the main tools for subject description used in library catalogue records, and are used for organizing internet resources in some web portals.

The best-known and most widely used examples are: the Dewey Decimal Classification (DDC), devised by the American librarian Melville Dewey and first published in 1876; the Universal Decimal Classification (UDC), first created in 1905 by the documentalists Paul Otlet and Henri La Fontaine, who have been mentioned in previous chapters, as an extension of Dewey's scheme; and the Library of Congress Classification (LCC), devised by that library, and begun in 1901. They have all undergone continual revision; Dewey, for example, is now in its 22nd edition. Even so, they show their origins: for example, Dewey's main structure shown above reflects both the 19th-century Western world, and the 'liberal arts' setting in which it was created. Although all widely used enumerative classifications have been updated in detail, major restructuring is unpopular because of the upheaval which would be caused to large libraries which use them for physical arrangement of material. For detailed treatment of the Dewey classification, see Bowman (2005), Chan (2007, Chapter 13) and Satija (2007).

Dewey, and particularly UDC, which was designed to deal better with detailed analysis of technical material than Dewey, have acquired a 'synthetic' or 'number building' facility in recent editions. Rather than list (enumerate) notations for every possible concept, notations can be built up. This capacity was present from the start, with tables of subdivisions, for types of materials, time periods and geographical areas, which could be linked to class numbers. This can be taken further by combining subjects. To give a simple example, to classify a book on 'agricultural research in Japan' in Dewey, we would first decide that the main concept was 'agriculture' to be qualified by the concepts of 'research' and 'Japan', the latter two taken from tables of general concepts, applicable in many cases. So, since agriculture has the notation 630, research is 072, and Japan is 052, we can construct the class number:

630.72052	agricultural research in Japan

UDC goes further than this, in that whole sections of the classification are 'synthetic', allowing class numbers to be built up as needed. In this way, the enumerative classifications are adopting some of the nature of faceted classifications, discussed later. For accounts of the UDC and its recent development, see McIlwaine (1997) and Slavic, Cordeiro and Riesthuis (2008).

The 'purest' example of a large enumerative classification is the Library of Congress Classification (LCC), which has limited synthetic capabilities. Despite this, and the fact that it is designed by and for one specific national library, it is increasingly popular in academic libraries worldwide, while the National Library of Medicine classification, effectively a sub-set of LCC, is widely used in healthcare libraries; in the UK, a local variant, the Wessex classification, is generally used. LCC provides very detailed, sometimes idiosyncratic, subject listings, based upon enumeration within 19 major classes, e.g. H for social sciences and R for medicine, giving entries such as:

HQ9261	reform and reclamation of adult prisoners
R601-602	food and food supply in relation to public health

For a detailed account of LCC, see Chan (2007, Chapter 14).

Finally, *analytico-synthetic* classifications, commonly termed faceted classifications, are designed to classify complex material, at a high level of subject specification. Terminology is grouped into related concepts by facet analysis (hence analytic), from which the classification for any item can be constructed (hence synthetic); these classifications can then cope with new concepts, in a way in which enumerative schemes cannot (although, as noted above, the enumerative schemes are gaining synthetic capabilities).

Faceted classifications were devised by the Indian librarian S. R. Ranganathan in the 1930s. Ranganathan's Colon Classification (named for the punctuation mark which characterizes its notation), the first universal classification of this type, has been little used outside the Indian subcontinent; for an overview of this scheme, see Satija and Singh (2010). Ranganathan's ideas led to the creation of many faceted classification schemes in specific subject areas, particularly in science, technology and social sciences. This type of scheme was very popular during the 1950s and 60s for classifying specialized or technical material, particularly in systems using techniques of mechanized documentation and, later, early applications of computers.

They rapidly lost popularity, however, because of their perceived complexity and difficulty of use – their notations are certainly unfriendly to the casual user, they are not well suited for physical arrangement of material, and they are relatively little used today. The only major scheme of this sort, apart from the Colon Classification, still under development is the second edition of the Bliss

Classification (BC2); the first edition was an enumerative scheme, and was entirely revised. This is regarded as an influential and theoretically sound scheme, but again has little practical use (Broughton, 2010).

We now move to the second major type of controlled vocabulary; alphabetic.

Alphabetic vocabularies: subject headings and thesauri

There are several kinds of alphabetic controlled vocabularies in which the terms – words or phrases – are arranged in alphabetical order. There are no well defined distinctions between them; in general, the distinction lies in how much information is provided about each term, including term definitions and inter-relations.

Keyword lists are the simplest form of alphabetic controlled vocabulary. Often they consist of nothing more than a list of 'approved terms'; sometimes they include synonyms. They are a simple and cheap form of terminology to develop and use, but are very limited in their usefulness, and applicable only to small files and unsophisticated users.

Subject headings are lists of terms – often quite lengthy to represent complex concepts – which are used for indexing for retrieval, and sometimes browsing. They will generally include synonyms, and sometimes hierarchical and 'SEE ALSO' relations. Complex subject heading lists can be very similar to thesauri. On the other hand, the simpler forms are little more than lists of keywords. They are most commonly used in situations where only one heading, or only a few, are added to each record, e.g. library databases and bibliographies.

The Library of Congress Subject Headings (LCSH) is the most widely known and widely used terminology of this sort; it provides subject indexing in many library databases, and increasingly in digital environments. A very large vocabulary, first published in 1914, it has over 270,000 terms, covering all subject areas. Examples of terms are:

Halloween cookery
Snails as carriers of disease
Overweight women in art
Electronic reserve collections in libraries
Virus diseases in children
Church work with the baby boom generation
Space flight on postage stamps.

These examples illustrate the way in which these headings link together several concepts. For a detailed overview of LCSH, see Broughton (2012), and for explanations of its applicability in digital resources see Walsh (2011) and Yi and Chan (2010).

Thesauri are of particular importance, as they are a sophisticated form of terminology, which can be very powerful in giving effective access to digital information; they are also the only retrieval terminology defined by national and international standards, though not all vocabularies called thesauri observe their prescriptions. For detailed coverage of thesauri, see Broughton (2006b) and – sound on the principles though dated in detail – Aitchison, Gilchrist and Bawden (2000).

Thesauri are listings of terms with inter-term relations shown. There are various relations which may be used, but the standard set (that is, literally, the set defined by the relevant international standards: ISO 2788 for monolingual thesauri) is:

SY	synonym
BT	broader term
NT	narrower term
RT	related term

Another useful relation is 'Top Term' or 'Heading Parent', identifying the hierarchy in which the term occurs.

One of the set of synonyms will be a 'preferred term', giving an unsymmetrical USE/USE FOR relation. All the other relations are generally symmetric. The standard also requires *scope notes*, notes giving definitions or explanations of terms, and/or prescribing their use in indexing. An example of a term in thesaurus form is shown in Figure 6.3 on the next page.

Thesauri have generally been used for both indexing and searching, but they may also be used in the form of an 'indexing thesaurus' (to provide extra terms to aid free-text searching) or a 'search thesaurus' (to suggest extra terms for a searcher to use in querying a full-text database).

The process of constructing a thesaurus involves an initial analysis of the subject area, in the same way that would be used to construct a faceted classification; indeed a thesaurus and a classification can be two 'faces' of the same scheme. More usually, the classification structure is used simply as a framework to derive the terms with their interrelationships that will form the thesaurus.

Having reviewed methods for describing resources and their subject content, we conclude this chapter by looking at two long-established ways of organizing and controlling information; the writing of abstracts and summaries of lengthy items, and the indexing of documents and collections.

Abstracting

An abstract is 'a brief but accurate representation of the contents of a document'

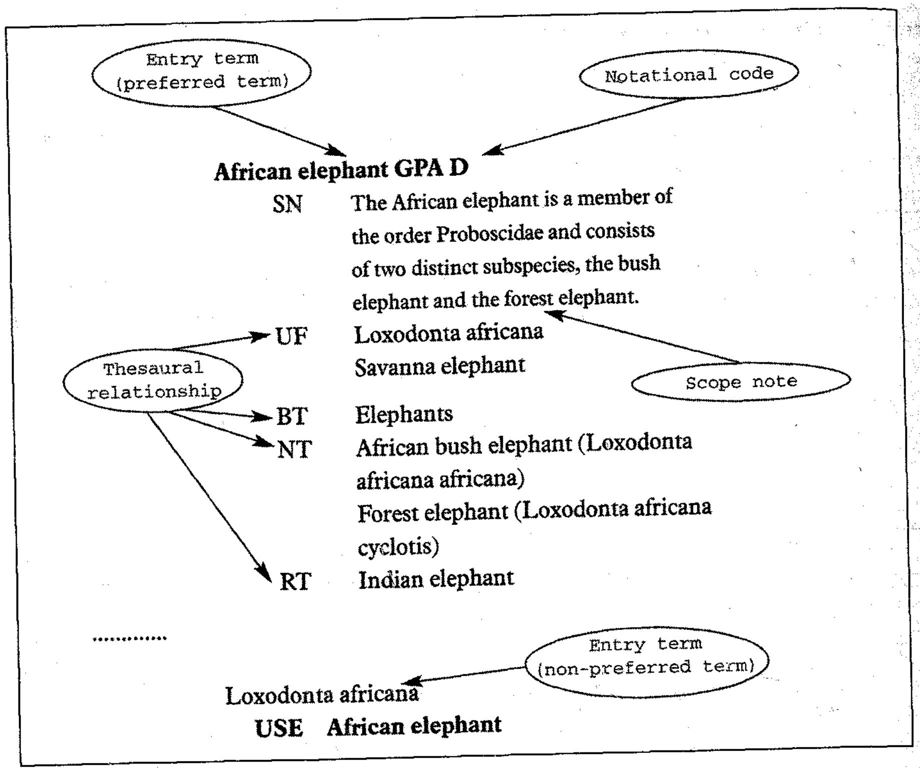


Figure 6.3 *Example of thesaurus term (reproduced from Broughton, 2006b)*

(Lancaster, 2003) or, more formally according to the relevant ISO standard 'an abbreviated, accurate representation of the contents of a document, without added interpretation or criticism and without distinction as to who wrote the abstract'.

Abstracting has a long history; summaries of articles have been used as means of keeping up with the scientific, medical and professional literature in particular for many years. Borko and Bernier (1975) suggest that its origins can be traced back to classical times, with the first recognizable abstracting journals, and sections for abstracts in general journals, dating from the 17th century, their use expanding greatly in the 19th century, together with the creation of subject-specialist abstracting and indexing services such as Index Medicus and Chemical Abstracts. For overviews, apart from the book by Borko and Bernier, see Chowdhury (2010, Chapter 8) and Koltay (2010); Alonso and Fernandez (2010) give a conceptual model for studying abstracts and abstracting.

Abstracts may vary in a number of respects, and so may be categorized differently. A fundamental difference is whether they are *informative*, intended to give sufficient information to be a replacement for the original, or *indicative*, giving just enough information for a reader to decide whether the item is of interest. They also vary:

- by length, from the verbose and 'literary' to the terse and 'telegraphic'
- by the extent of criticism and interpretation of the original
- by the extent to which they are 'targeted' or 'slanted' to a particular interest, or type of user, as against attempting to be balanced and objective.

The writing of abstracts is governed by international standard: ISO 214 (1976) 'Abstracts for publications and documentation'. Publishers and database producers also have *de facto* standards; see, for example, Montesi and Owen (2007). There is a general tendency for abstracts to become more 'structured', with a consistent set of elements which are present. There still remains debate about the most effective ways to make abstracts useful; see, for example, Hartley and Betts (2008, 2009), Zhang and Liu (2011) and Ripple et al. (2011). This is of particular importance as more reliance is placed on abstracts because of information overload (Nicholas, Huntington and Jamali, 2007). This is potentially troubling, as studies have shown that typically 20% of abstracts contain significant inaccuracies – usually presenting the subject matter of the main document in an unreasonably positive light.

Automatic abstracting has been a goal for many years, since the first research on the topic, carried out by H. P. Luhn in the late 1950s. Much effort has been expended since then on the design of systems, using a variety of approaches to automatic abstracting and summarizing: see Chowdhury (2010) for an overview. Nonetheless, most practical abstracting within information systems and services is still largely an intellectual task.

Indexing and tagging

An index is usually understood as a systematic arrangement of entries designed to enable users to locate information in a document, or in a collection of documents. The process of producing such an index is 'indexing', and those who do it are 'indexers'; it is governed by international standards ISO 5963 (1985) 'Examining documents, determining their subjects and selecting index terms' and ISO 999 (1996) 'Guidelines for the content, organization and presentation of indexes'. The classic text on the indexing process is Lancaster (2003); for a review of the theoretical and historical background, see Weinberg (2009).

It is usually taken to mean the assignment of a number of terms from an alphabetic vocabulary, by contrast with classifying, where a systematic vocabulary is used. If no controlled vocabulary is used, the process is called 'free term indexing'. While this is likely to capture very precise and up-to-date terminology, the use of isolated keywords means that no semantic relations between terms can be made visible. Free keywording has attracted interest recently, as 'folksonomy' or 'social tagging', whereby internet materials, such as web pages, photographs, videos and catalogue records for books, are freely indexed by users (Ding et al., 2009; Mai, 2011; Park, 2011; Voorbij, 2012). The long-term usefulness of this, and how it might be combined with, or complement, traditional intellectual indexing, is unclear. One response from information practitioners has been to look for ways to combine tagging with controlled vocabulary structures; see, for example, Šauperl (2010) for the UDC, and Yi and Chan (2009) for LCSH .

There are many types of index: 'back of the book' indexes; indexes of the content of a journal issue or volume; cumulative indexes to journals, newspaper and magazines; database indexes; and indexes to pages on internet or intranet sites. An index may be to a *collection of items*, e.g. indexing a database, an intranet, or an Internet directory, or to a *single item*, e.g. indexing the contexts of a book, or a report, or a volume of a journal. In the case of collections, the index directs users to a particular item within the collection. In the single item case, the index 'points to' a page or a section within the item.

The basic principles in the production of any index – the 'indexing process' – are largely the same, regardless of the type of index. Indexing is a process of firstly deciding what the item being indexed is about ('content analysis') and then deciding how best to represent this 'aboutness' ('term selection'). The indexer must identify abstract concepts, and then match these with appropriate terms. This may be contrasted with the simple automatic indexing of full-text, in which terms are selected from the document on the basis of concordance (all words are chosen), statistical analysis (terms which occur relatively frequently are chosen) or positional analysis (terms in the title, abstract, first paragraph, etc. are chosen) This latter approach can be done perfectly well by machine, since it deals simply with text strings, with no identification of underlying concepts. It is this concept analysis which makes indexing an intellectual process. The justification for continuing to use human, intellectual indexing, rather than the much cheaper automatic indexing, is that human analysis gives an index which is more useful and usable. Though not particularly consistent: many studies have shown that the best degree of consistency which can be obtained between human indexers does not exceed 50%.

Intellectual indexing is claimed to have a number of significant advantages over full-text searching, or production of concordances by computer – which amounts to the same thing – in addition to concept analysis. The human indexer can deal readily with:

homographs – words which are spelt the same, but have different meanings;
synonyms – different words with the same meaning;
inferences – where the main subject word is implied, but never used;
and with the difference between significant and trivial 'passing' mentions of a word.

Automatic indexing systems, although becoming much more capable, have problems with all these aspects. A variety of software systems is available, to produce indexes automatically or to aid the human indexer, such as Cindex, Macrex and Sky Index (Coates, 2009).

Intellectual indexing of a particular document can be regarded as a three-stage process:

- understanding of the content: the indexer decides what the document is 'about'
- analysis of the content: the indexer decides which concepts are to be indexed, and to what depth
- translation of concepts into indexing terms: the indexer chooses the most appropriate terms from the alphabetic vocabulary being used (e.g. a thesaurus or list of subject headings).

The two generally accepted principles of indexing are:

- include all concepts thought to be of interest to users of the index
- index at the most specific level that the indexing vocabulary allows.

These are rather general ideas; the indexer has to interpret them. The two main criteria for indexing are:

- exhaustivity: the extent to which all possible concepts are included
- depth: the degree of specificity with which concepts are described.

Indexing which is both exhaustive and deep is usually regarded as the 'gold standard', but the other three possibilities (deep but not exhaustive, etc.) may be appropriate in different circumstances.

Summary

Tools for information organization have changed slowly; some of these used today, more than a century after they were first introduced, would seem familiar to Melville Dewey and Anthony Panizzi. The dramatic changes in the technical means by which documents are created and disseminated has not been matched by the intellectual means by which they are managed. Whether automated methods will be devised to do the same job as current tools, and whether folksonomies and the like will prove effective in the long term, remains to be seen.

- Information organization remains at the heart of information science, its importance enhanced in new information environments
- New forms of descriptive metadata have emerged to deal with new forms of document, while subject description has been little altered.
- The right balance between expert human input, automated processes, and social tagging has yet to be established
- Long-established theories and concepts remain important, despite technical advances

Key readings

Elaine Svenonious, *The intellectual foundation of information organization*, Cambridge MA: MIT Press, 2000.

[A thorough treatment of basic issues.]

Grigoris Antoniou and Frank van Harmelen (2008) *A semantic web primer* (2nd edn), Cambridge MA: MIT Press.

[Accessible analysis of the new forms of information organization.]

References

Aitchison, J., Gilchrist, A. and Bawden, D. (2000) *Thesaurus construction and use: a practical manual* (4th edition), London: Aslib.

Allinson, J. (2012) OpenART: open metadata for art research at the Tate, *Bulletin of the American Society for Information Science and Technology*, 38(3), 43–8.

Alonso, M. I. and Fernandez, L. M. (2010) Perspectives of studies on document abstracting: towards an integrated view of models and theoretical approaches, *Journal of Documentation*, 66(4), 563–84.

Anhalt, J. and Stewart, R. A. (2012) RDA simplified, *Cataloging and Classification Quarterly*, 50(1), 33–42.

Antoniou, G. and van Harmelen, F. (2008) *A semantic web primer* (2nd edn), Cambridge MA: MIT Press.

Ayre, C., Smith, I. A. and Cleeve, M. (2006) Electronic library glossaries: jargonbusting essentials or wasted resource?, *Electronic Library*, 24(2), 126–34.

Beghtol, C. (2010) Classification theory, *Encyclopedia of Library and Information Sciences* (3rd edn), Abingdon: Taylor & Francis, 1045–60.

Borko, H. and Bernier, C. L. (1975) *Abstracting concepts and methods*, New York NY: Academic Press.

Bowker, G. C. (2005) *Memory practices in the sciences*, Cambridge MA: MIT Press.

Bowker, G. C. and Star, S. L. (2000) *Sorting things out: classification and its consequences*, Cambridge MA: MIT Press.

Bowman, J. (2003) *Essential cataloguing*, London: Facet Publishing.

Bowman, J. (2005) *Essential Dewey*, London: Facet Publishing.

Bowman, J. H. (2006) The development of description in cataloguing prior to ISBD, *Aslib Proceedings*, 58(1/2), 34–48.

Broughton, V. (2005) *Essential classification*, London: Facet Publishing.

Broughton, V. (2006a) The need for a faceted classification as the basis of all methods of information retrieval, *Aslib Proceedings*, 58(1), 49–72.

Broughton, V. (2006b) *Essential thesaurus construction*, London: Facet Publishing.

Broughton, V. (2010) *Bliss Bibliographic Classification Second Edition, Encyclopedia of Library and Information Sciences*, Abingdon: Taylor & Francis, 1:1, 650–9.

Broughton, V. (2012) *Essential Library of Congress Subject Headings*, London: Facet Publishing.

Buckland, M.K. (2008) Reference library service in the digital environment, *Library and Information Science Research*, 30(2), 81–5.

Cassell, K. A. and Hiremath, U. (2011) *Reference services in the 21st century* (2nd edn revised), London: Facet Publishing.

Chan, L. M. (2007) *Cataloguing and classification: an introduction* (3rd edn), Lanham MD: Scarecrow Press.

Chowdhury, G. G. (2010) *Introduction to modern information retrieval* (3rd edn), London: Facet Publishing.

Chowdhury, G. G. and Chowdhury, S. (2007) *Organizing information: from the shelf to the web*, London: Facet Publishing.

Coates, S. (2009) Software solutions, *Indexer*, 27(4), 168–72.

Coyle, K. and Hillman, D. (2007) Resource Description and Access (RDA). Cataloguing Rules for the 20th Century [sic], *D-LIB Magazine*, 13(1/2), Jan/Feb 2007, available from: http://www.dlib.org/dlib/january07/coyle/01coyle.html.

Dimic, B., Milsavljevic, B. and Surla, D. (2010) XML schema for UNIMARC and MARC21, *Electronic Library*, 28(2), 245–262.

Ding, Y., Jacob, E. K., Zhang, Z., Foo, S., Yan, E., George, N. L. and Guo, L. (2009) Perspectives on social tagging, *Journal of the American Society for Information Science and Technology*, 60(12), 2388–2401.

Guerrini, M. (2009) In praise of the unfinished: the IFLA Statement of International Cataloguing Principles 2009, *Cataloguing and Classification Quarterly*, 47(8), 722–40.

Hartley, J. and Betts, L. (2008) Revising and polishing a structured abstract: is it worth the time and effort? *Journal of the American Society for Information Science and Technology*, 59(12), 1870–7.

Hartley, J. and Betts, L. (2009) Common weaknesses in traditional abstracts in the social sciences, *Journal of the American Society for Information Science and Technology*, 60(10), 2010–18.

Haynes, D. (2004) *Metadata for information management and retrieval*, London: Facet Publishing.

Heidorn, P. B. (2011) Biodiversity informatics, *Bulletin of the American Society for Information Science and Technology*, 37(6), 38–44.

Hider, P. (2012) *Information resource description: creating and managing metadata*, London: Facet Publishing.

Hitchens, H. (2005) *Dr Johnson's Dictionary – the extraordinary story of the book that defined the world*, London: John Murray.

Hjørland, B. (2003) Fundamentals of knowledge organization, *Knowledge Organization*, 30(2), 87–111 .

Hjørland, B. (2008a) What is knowledge organization (KO)?, *Knowledge Organization*, 35(2/3), 86–101.

Hjørland, B. (2008b) Core classification theory: a reply to Szostak, *Journal of Documentation*, 64(3), 333–42.

Hjørland, B. (2009) Concept theory, *Journal of the American Society for Information Science and Technology*, 60(8), 1519–36.

Hjørland, B. (2011) The importance of theories of knowledge: indexing and information retrieval as an example, *Journal of the American Society for Information Science and Technology*, 62(1), 72–7.

Hjørland, B. and Pedersen, K. N. (2005) A substantive theory of classification for information retrieval, *Journal of Documentation*, 61(5), 582–97.

Hjørland, B. and Shaw, R. (2010) Concepts: classes and colligation, *Bulletin of the American Society for Information Science and Technology*, 36(3), 2–4, available from http://www.asis.org/Bulletin/Feb10/Bulletin_FebMar10_Final.pdf.

Hüllen, W. (2004) *A history of Roget's Thesaurus: origins, development and design*, Oxford: Oxford University Press.

Hunter, E. J. (2009) *Classification made simple* (3rd edn), Aldershot: Ashgate.

IFLA (1998) Functional Requirements for Bibliographic Records, produced by the IFLA, Study Group on the Functional Requirements for Bibliographic Records, Munich: K. G. Saur: available from http://www.ifla.org/VII/s13/frbr/frbr.pdf.

Koltay, T. (2010) *Abstracts and abstracting: a genre and skills for the 21st century*, Oxford: Chandos.

La Barre, K. (2010) Facet analysis, *Annual Review of Information Science and Technology*, 44, 243–86.

Lambe, P. (2007) *Organizing knowledge: taxonomies, knowledge and organisational effectiveness*, Oxford: Chandos.

Lancaster, F. W. (2003) *Indexing and abstracting in theory and practice* (3rd edn), London: Facet Publishing .

Langridge, D. W. (1992) *Classification: its kinds, systems, elements, and applications*, London: Bowker-Saur.

Leigh, G. J. (2011) *Principles of Chemical Nomenclature 2011: a guide to IUPAC recommendations*, London: Royal Society of Chemistry.

Mai, J-E. (2011) Folksonomies and the new order: authority in the digital disorder, *Knowledge Organization*, 38(2), 114–22.

McIlwaine, I. C. (1997) The Universal Decimal Classification: some factors concerning its origins, development and influence, *Journal of the American Society for Information Science*, 48(4), 331–9.

Mersky, R. M. (2004) The evolution and impact of legal dictionaries, *Legal Reference Services Quarterly*, 23(1), 19–35.

Miller, S. J. (2011) *Metadata for digital collections: a how-to-do-it manual*, London: Facet Publishing.

Montesi, M. and Owen, J. M. (2007) Revision of author abstracts: how it is carried out by LISA editors, *Aslib Proceedings*, 5991, 26–45.

Mugglestone, L. (2005) *Lost for words: the hidden history of the Oxford English Dictionary*, New Haven CT: Yale University Press.

Mugglestone, L. (2011) *Dictionaries: a very short introduction*, Oxford: Oxford University Press.

Nicholas, D., Huntington, P. and Jamali, H. R. (2007) The use, users, and role of abstracts in the digital scholarly environment. *Journal of Academic Librarianship*, 33(4), 446–53.

Oliver, C. (2010) *Introducing RDA: a guide to the basics*, London: Facet Publishing.

Park, H. (2011) A conceptual framework to study folksonomic interaction, *Knowledge organization*, 38(6), 515–29.

Patton, G. E. (ed.) (2009) *Functional requirements for authority data: a conceptual model*, Munich: K. G. Saur.

Pisanski, J. and Žumer, M. (2010) Mental models of the bibliographic universe. Part 1: mental models of descriptions, *Journal of Documentation*, 66(5), 643–67.

Ripple, A. M., Mork, J. G., Knecht, L. S. and Humphries, B. L. (2011) A retrospective cohort study of structured abstracts in MEDLINE, 1992–2006, *Journal of the Medical Library Association*, 99(2), 160–2.

Robinson, L. (2010) *Understanding healthcare information*, London: Facet Publishing.

Salaba, A., Zeng, M. L. and Žumer, M. (eds) (2011) *Functional requirements for subject authority data (FRSAD): a conceptual model*, Munich: de Gruyter.

Satija, M. P. (2007) *The theory and practice of the Dewey Decimal Classification system*, Oxford: Chandos.

Satija, M.P. and Singh, J. (2010) Colon Classification (CC), *Encyclopedia of Library and Information Sciences* (3rd edn), Abingdon: Taylor & Francis, 1:1, 1158–68.

Šauperl, A. (2010) UDC and folksonomies, *Knowledge Organization*, 37(4), 307–17.

Slavic, A. (2011) Classification revisited: a web of knowledge, in Foster, A. and Rafferty, P. (eds), *Innovations in information retrieval*, London: Facet Publishing, pp 23–48.

Slavic, A., Cordeiro, M. I. and Riesthuis, G. (2008) Maintenance of the Universal

Decimal Classification: overview of the past and preparation for the future, *International Cataloguing and Bibliographical Control*, 37(2), 23–9.

Smiraglia, R. P. (2001) *The nature of a work: implications for the organization of knowledge*, Lanham MD: Scarecrow Press.

Smith, B., Ceusters, W., Klagges, B., Köhler, J., Kumar, A., Lomax, J., Mungall, C., Neuhaus, F., Rector, A. L. and Rosse, C. (2005) Relations in biomedical ontologies, *Genone Biology*, 6(5):r46 [online] available at http://genomebiology.com/content/pdf/gb-2005-6-5-r46.pdf.

Svenonious, E. (2000) *The intellectual foundation of information organization*, Cambridge MA: MIT Press.

Tackabery, M. K. (2005) Defining glossaries, *Technical Communication*, 52(4), 427–33.

Taylor, A. G. (2004) *The organization of information* (2nd edn), Santa Barbara CA: Libraries Unlimited.

Tennis, J. T. (2008) Epistemology, theory and methodology in knowledge organization: towards a classification, metatheory and research framework, *Knowledge Organization*, 35(2/3), 102–12.

Tillett, B. B. and Cristán, A. L (2009) *IFLA cataloguing principles: the statement of international cataloguing principles (ICP) and its glossary in 20 languages*, Munich: K. G. Saur. Also available online at http://www.ifla.org/publications/statement-of-international-cataloguing-principles.

Voorbij, H. (2012) The value of LibraryThing tags for academic libraries, *Online Information Review*, 36(2) [online EarlyCite file.].

Walsh, J. (2011) The use of Library of Congress Subject Headings in digital collections, *Library Review*, 60(4), 328–43.

Weinberg, B. H. (2009) Indexing: history and theory, *Encyclopedia of Library and Information Sciences* (3rd edn), Abingdon: Taylor & Francis, 1:1, 2277–90.

Welsh, A. and Batley, S. (2012) *Practical cataloguing: AACR, RDA and MARC21*, London: Facet Publishing.

Yi, K. and Chan, M. L. (2009) Linking folksonomy to Library of Congress subject headings: an exploratory study, *Journal of Documentation*, 65(6), 872–900.

Yi, K. and Chan, L. M. (2010) Revisiting the syntactical and structural analysis of Library of Congress Subject Headings for the digital environment, *Journal of the American Society for Information Science and Technology*, 61(4), 677–87.

Zeng, M. L. and Qin, J. (2008) *Metadata*, London: Facet Publishing.

Zhang, C.and Liu, X. (2011) Review of James Hartley's research on structured abstracts, *Journal of Information Science*, 37(6), 570–6.

Zhang, Y. and Salaba, A. (2009) *Implementing FRBR in libraries: key issues and future directions*, New York NY: Neal-Schuman.

Zins, C. and Santos, P. (2011) Mapping the knowledge covered by library classification schemes, *Journal of the American Society for Information Science and Technology*, 62(5), 877–901.

CHAPTER 7

Information technologies: creation, dissemination and retrieval

> The change from atoms to bits is irrevocable and unstoppable . . . Computing is not about computers any more. It is about living.
>
> Nicholas Negroponte (1995, 4 and 6)

> As more information is represented digitally, human-computer interaction (HCI) broadly defined, becomes more central to information science.
>
> Jonathan Grudin (2011, 369)

Introduction

In this chapter, we will give an overview of the information technologies which underlie the information sciences, and some of the more important application areas. This is obviously a very wide area, and whole books are written about many of the topics within it. We shall attempt no more than to mention and briefly discuss each of the topics within this area and give references to where more details can be found. Our aim is simply to give an overall picture of what technologies are important to the information scientist and how they relate to one another. Many readers will be familiar with much of this material and we ask them to consider this a refresher course.

We will look initially at the nature of technology generally, and information technology in particular, setting the scene for what follows. We will then examine the nature of digital computers and their software systems, the networks which connect them, some of the new physical forms they are taking, and some ideas of the future of computing. We will cover the ways in which people interact with computers, and the ways in which IT systems are envisaged and designed. Finally we consider some of the important applications – particularly information retrieval and digital libraries – showing how they can be understood as following and facilitating the communication chain.

What are information technologies?

'Technology', from the Greek *techné*, meaning art, skill or craft, is usually taken to mean the understanding of how to use tools, in the broadest sense of the word.

The term *information technology* was first used in the 1950s, to describe the application of mechanized documentation and the new digital computers, but it came to be widely used in the 1980s to describe the much wider spread use of computers, particularly the first personal computers, and of the computer networks which pre-dated the internet. Early in that decade, Peter Zorkoczy (1982, 3), in one of the earliest books devoted to the subject, commented that the phrase was 'a relatively recent and perhaps not particularly well chosen addition to the English language'. Pointing to the various ways in which the term was understood, he declined to give an exact definition; we will follow his example.

The concept of information technology is usually associated with computers and networks. But, in a wider sense stemming from the original meaning of the word, the technologies of information include all the tools and machines which have been used to assist the creation and dissemination of information throughout history, as discussed in Chapter 2; from ink and paper, through printing, to mechanized documentation technologies and the photocopier. Ben Shneiderman (2003) gives a thoughtful mediation on the nature of information technology, and its empowering effects, while Nicholas Negroponte (1995) provides an early, and very prescient, account of the impacts of IT; Markus Krajewski (2011) examines the idea of card index files as a 'universal paper machine', the forerunner of the computer.

While not ignoring the past and present significance of these, we will focus here on digital technologies. In doing so, we take the view expressed by Paul Gilster, whom we will meet again in Chapter 13, to the effect that all information today is digital, has been digital, or may be digital. And, pragmatically, digital technologies are more complex and rapidly changing than others.

This last point is summed up well by Moore's Law, which states that the number of components which can be placed inexpensively into the integrated circuits which are the basis of all modern digital devices roughly doubles every two years. This means that processing speed, storage capacity, and other metrics of computer power also increase at the same rate.

Impressive though this is, advances in information technology are not due to this alone. Largely as a result of these advances, digital devices have become more interconnected with each other, more integrated into other sorts of equipment and product, much smaller, and more pervasive, affecting all aspects of life and work which have any information component. With this background, we will now look in outline at the digital technologies which have affected all aspects of the communication chain over the past decades.

Digital technologies

We will describe these aspects only in outline; see Ince (2011) and White and Downs (2007) for more detailed but still accessible accounts, and for a readable account of the historical development of the digital computer see Hally (2005).

Any digital device represents data in the form of *binary digits* or *bits*. Patterns of bits may represent data or instructions. A collection of eight bits is known as a *byte*. Quantities of data are represented as multiples of bytes, for example:

kilobyte	one thousand	(10^3) bytes
megabyte	one million	(10^6) bytes
gigabyte	one billion	(10^9) bytes

There is an older convention, based on binary notation which was rather different. A kilobyte would be defined as 10^2bytes, i.e. 1024 bytes, rather then 1000. A gigabyte, under this understanding, is actually 1,073,741,824 bytes. These variants have now been renamed kibibytes, mebibytes and gibibytes by the standards authorities, but the ambiguity persists.

Any character or symbol may be represented in binary notation, but this requires an agreed coding. The most widely used code since the beginning of the computer age has been the American National Standards Institute's ASCII (American Standard Code for Information Interchange) code, but it is limited to the Latin alphabet, Arabic numerals and a few other symbols. It is being supplanted by Unicode, which can handle a wider variety of symbols and scripts. It can do this by having a long coding string, up to 32 bits, while ASCII is restricted to 7 bits; the more bits in the code, the more different symbols can be coded. Codes provide arbitrary representations of characters; for example, the ASCII code for the letter L is 1001100.

The basic architecture of the digital computer has not changed since it was set out by John von Neumann, shown in Figure 7.1 on the next page with an early working computer, in 1945. A Hungarian-born mathematician and physicist, von Neumann spent most of his life in the USA, working on a variety of topics, including nuclear physics, quantum mechanics, game theory, information theory and – not least – the fundamentals of computing. His design, which gave the first formal description of a single-memory stored-program computer, is shown, in a modernized form, in Figure 7.2 (on page 135). This is generally referred to as the von Neumann architecture, although he clearly drew from his collaborations with other US computer pioneers, such as Presper Eckert and John Mauchly.

This architecture is general-purpose, in the sense that it can run a variety of programs. This distinguishes it from special-purpose digital computers which carry out only one task; there, are several such computers, for example, in the

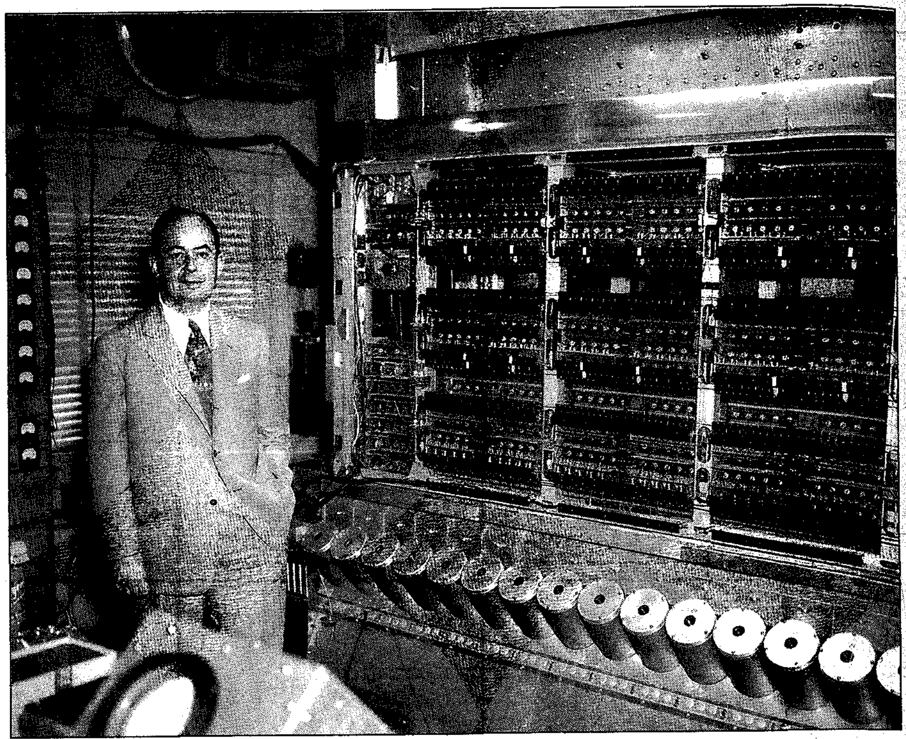


Figure 7.1 *John von Neumann with the computer of the Institute of Advanced Study (Alan Richards photographer. From The Shelby White and Leon Levy Archives Center, Institute for Advanced Study, Princeton, NJ, USA)*

appliances in most kitchens in the developed world and several in any modern car. A von Neumann machine loads and runs programs as necessary, to accomplish very different tasks. Ince (2011, 6–7) expresses this in words, to give a working definition of a computer:

> A computer contains one or more processors which operate on data. The processor(s) are connected to data storage. The intentions of a human operator are conveyed to the computer via a number of input devices. The result of any computation carried out by the processor(s) will be shown on a number of display devices.

The heart of the computer, the *processor*, often referred to as the *central processing unit (CPU)*, carries out a set of very basic arithmetic and logical operations with instructions and data pulled in from the *memory*, also referred to as *main* or *working* memory. This sequence is referred to as the *fetch-execute cycle*. Two components of the processor are sometimes distinguished: an *arithmetic and logic unit*, which carries out the operations, and a *control unit*, which governs the operations of the cycle.

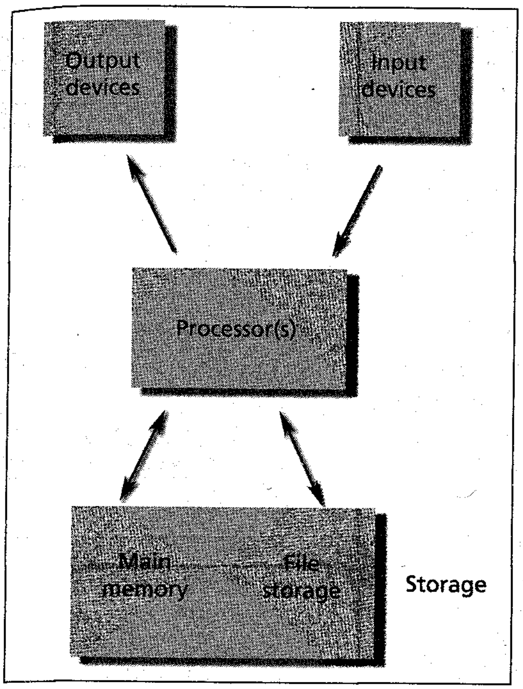


Figure 7.2
Von Neumann architecture (Reproduced from Ince, 2011, by permission of Oxford University Press)

While programs and data are being used, they are kept in the memory. While they are not being used, they are stored long-term in the *file storage*. Items in memory are accessible much more rapidly than items in file storage. Memory is much more expensive, so computers have much more file storage than memory; the computers on which this book was written have 1 and 2 gigabytes of memory and 75 and 250 gigabytes of file storage respectively. Although these are much larger than the memory and file storage on computers of past years, the principle remains the same. For example, the IBM XT personal computer, introduced in 1983, had 256 kilobytes of memory and 10 megabytes of file storage.

Data comes into the computer through its *input devices* and is sent into the outside world through the *output devices*. The components are linked together through circuits usually denoted as a *bus*, sometimes referred to more specifically as a *data bus* or an *address bus*.

All of these components have undergone considerable change since the first computers were designed. Processor design has gone through three main technological stages. The so-called 'first generation' of computers used valves as processor components, and the 'second generation' used transistors. Computers of the 'third generation', including all present-day computers, use circuits on 'silicon chips', using methods of very-large-scale integration (VLSI), which allows millions of components to be placed on a single computer chip; see Ince (2011) for a readable introduction to this technology. The tangible result of this has been a great increase in processing speed, typically measured as the number of instructions per second; the computer on which these words are being typed has a processor speed of 2 Gigahertz, meaning two billion instructions per second.

I – III. Generace PC

The storage elements of the computer's memory comprise regular arrays of silicon-based units, each holding a bit of data, and created using the same VLSI methods as processors. Earlier generations of computers used so-called 'core storage', with data held on tiny magnetized elements arranged in a three-dimensional lattice; again each element held one bit of data.

File storage, holding data and programs that are not needed immediately, has always used magnetic media, which can hold large volumes of data cheaply, provided quick access is not required. Earlier generations of computers used a variety of tapes, drums and disks; current computers use so-called *hard disks*, rapidly spinning magnetizable disks, with moveable arms with *read/write heads*, to read data from or write data to any area of the disk. Regardless of the technology, each area holds one bit of information, according to its magnetic state.

Input devices fall into three categories: those which take input interactively from the user; those which accept data from other digital sources; and those which convert paper data into digital form. The first category comprises the venerable QWERTY keyboard, originally developed for typewriters in the 1870s, together with more recently developed devices: the mouse and other pointing devices, and the touch-screen. The second category comprises the silicon-memory data stick, replacing the various forms of portable magnetic 'floppy disks' used previously, and the ports and circuits by which the computer communicates with networked resources. The third category is that of the scanner, digitizing print and images from paper sources.

Output devices are similarly categorized in the same three ways. There are the display screens, which allow user interaction through visual and sound output; those which output digital data – data sticks and network circuits and ports; and those which print paper output; typically now laser or inkjet printers for personal and office use, and commercial digital printers. These latter have revolutionized the production of printed documents, as we shall discuss further in Chapter 10, for two main reasons. They make the unit cost of the printing of one document the same as that of many, allowing, most notably, print-on-demand books; for the impact of such devices in a library setting, see Arlitsch (2011). And they allow text and images to be handled in an integrated way, since both are represented as binary data patterns, and the basic printing unit for each is the pixel, the smallest point which can be displayed on a screen or printed on a page. This contrasts with early forms of printing, in which they had to be treated in a different way; older books usually have images on separate pages from print. Computerized phototypesetting began to be employed in the 1960s, with full digital printing in the 1980s; see Cope and Phillips (2006) and Twyman (1998) for more details, and Chapter 10 for more on the consequences.

Having outlined the nature of the isolated digital computer, we will now consider the networks which connect them together.

Networks

No computer today lives an isolated life. All are connected via some form of network to others, to enable communication, and information access and sharing. Since the 1990s the internet and the world wide web have become ubiquitous, such that it has become difficult to think of information technologies without these at centre stage. We will look in outline at some network concepts and examples, then moving on consider two recent developments: the grid and the cloud. For more details, presented in an accessible way, see Ince (2011), Davis and Shaw (2011, Chapter 6), Derfler and Freed (2004), and Gralla (2006).

The growth of networked computing has been driven by three factors: communications technology, software and standards. In terms of technology, older forms of network, based on co-axial cables originating in the 1880s and used for telegraph and telephone systems, have been succeeded by fibre-optic cables, and various forms of wireless transmission. These allow much faster transmission speeds and greater information carrying capacity; such systems are described loosely as *broadband*. Software systems have improved the efficiency and reliability of transmission greatly: an important example is *packet switching*, by which messages may be split up and their constituents sent by the fastest available route, being recombined before delivery.

Standards are fundamentally important, so that different forms of network, and different kinds of computers connected to them, can communicate effectively. The internet, originating in the 1960s in networks built for defence research in the USA, is a worldwide network of networks, integrated by common standards: the internet control protocols, commonly known as TCP/IP (Transmission Control Protocol/Internet Protocol). It has been estimated that in 2011 over two billion people, one third of the world's population, were users of the internet.

The world wide web, often spoken of as if it were synonymous with the internet, is in fact a major internet application. A system for allowing access to interlinked hypertext documents stored on networked computers, it originated in the work of Sir Tim Berners-Lee at the CERN, the European nuclear research establishment. Berners-Lee introduced the idea in 1989 in an internal memorandum with the modest title 'Information management: a proposal'. At the time of writing (December 2011), the original document was available at http://www.w3.org/History/1989/proposal.html. For Berners-Lee's own account of the origins and early nature of the web, see Berners-Lee (1999).

The web is based on so-called client-server architecture: a web browser, the client, on the user's computer accesses the website on the remote server computer, through the internet, and downloads the required web page. This relies on a number of standards. For example, web pages must be created using a mark-up language, typically HTML, sometimes referred to as the *lingua franca* of the internet, and be identified by a Uniform Resource Identifier, commonly referred to as a Uniform Resource Locator or URL; a Hypertext Transfer Protocol (HTTP) enables communication between client and server. Use of the web parallels that of the internet itself, with over two billion users estimated in 2011.

The internet and web, despite their great reach and influence, are by no means the only significant computer networks. Many private, local and regional networks exist, although increasingly they are using the standards of the internet and web, for compatibility.

Beyond the current network arrangement, we can see the development of new ways of connecting computers together.

Grid computing is a term used to describe a number of computers linked together by internet connections, and sometimes by high-speed network connections, so that they can work together, giving, in combination, the power of a much larger single machine; see Ince (2011) for a more detailed account, and Town and Harrison (2010) for an information retrieval example.

The *cloud* is a concept which takes networking to its ultimate extent. Rather than software and files of information being stored on individual computers, they are stored remotely, 'in the cloud' – in practice in large 'server farms' – and accessed via networks when they are needed. This is facilitated by 'cloud services', such as Apple's iCloud and Googledocs+, which control the servers and network access. Moulaison and Corrado (2011) give a good description of this technology, with case studies of particular relevance to information practitioners. We will return to the cloud as a possible future for information provision in the final chapter.

By comparison with earlier forms of computer, today's are very much more reliant on, and empowered by, network connectivity. But they are different in another way; their shapes and sizes.

Mobile and pervasive

The very first computers occupied a building. Subsequent generations occupied a large room, a small room and a desktop. The trend has continued, to give ever smaller and more mobile computing devices, exemplified by today's laptops, notebooks, tablet computers and smartphones. For examples relevant to library and information services, and issues associated with their use, see Ally and Needham (2012) and Wisniewski (2011).

The trend continues, with computing power appearing in, and being applied to, objects which would never have previously been associated with computers; a trend known as *pervasive* computing. We will give just two examples.

1 Radio Frequency Identification (RFID) tags are very small electronic devices, which may be attached to any kind of object, and which emit radio signals; these can be read by a receiver to locate and identify the object. One obvious example is insertion in library books, where they can aid location, circulation etc. (Walsh, 2011; Palmer, 2009; Zimerman, 2011).
2 QR (Quick Response) codes are *matrix barcodes*, black and white patterns in a square format. Originally they were designed for tracking items in manufacturing or retail settings, but they have gained much wider use since many smartphones gained apps for scanning them and using the encoded information, typically displaying text or opening a webpage. They are used in museums, galleries, libraries and information centres to provide information about collection items, to provide additional information about resources and services, to offer guides, maps, audio tours and self-service hints and tips, and for marketing and promotion; for detailed and varied examples, see Ekart (2011), Walsh (2011), Whitchurch (2011) and Hoy (2011).

So far, we have thought mainly about the physical devices of information technology: the *hardware*. To make a computer do any useful tasks requires instructions, in the form of programs: the *software*.

Software

Computer users will, for the most part, interact with two forms of software: *operating systems* and *applications*.

The operating system, specific to a particular type of computer and installed on each machine, controls the hardware, and runs applications. Examples are versions of Microsoft's Windows, Apple's OS X, and Linux.

Applications software typically comes in the form of packages, for a specific purpose. Applications may either be installed on individual computers, as for example word processors, spreadsheets and web browsers typically are, or may be accessed on the web; this is usual for search engines, library management systems, social media, and the like. As we have seen, there is an increasing trend for virtually all software to be accessible from the cloud, rather then installed on individual machines.

Applications software, from the users' perspectives, typically requires the computer to carry out tasks defined at a fairly high level: search the bibliographic database for authors with this name; calculate the mean of this column of figures; insert this image into the blog post; and so on. These must be translated into a series of very specific low-level commands to be carried out by the processor. This is achieved by software features operating 'behind' the interface of the application, usually unknown to the user.

All software must be written in some kind of programming language, though this will usually be invisible to the user, who will have no reason to know what language any particular application is written in. At the risk of over-simplification, we can say there are four kinds of software language:

1 *Low-level* programming languages, also referred to as *assembly* languages or *machine code*, encode instructions at the very detailed level of processor operations; these languages are therefore necessary specific to a particular type of computer. This kind of programming is complex and difficult, but results in very efficient operation; it is reserved for situations where reliably fast processing is essential.
2 *High-level* programming languages express instructions in terms closer to user intentions, and are converted by other software systems, generally termed *compilers*, into processor instructions; programs are therefore much easier to write and to understand, and the languages can be used on different types of computer. There have been many such languages, some general purpose and some aimed a particular kind of application. The first examples were developed in the 1950s, and two of the earliest are still in use: Fortran, still a language of choice for scientific and engineering applications, and COBOL, still used in many 'legacy' systems for business applications. Currently popular are Java, C++, and C# (pronounced C sharp).
3 *Scripting* languages are a particularly significant form of high-level language, designed to support interaction, particularly with web resources; for example, to update a web page in response to a user's input, rather than reloading the page each time, or to create a *mashup*, an integration of data from several web sources. Examples are Javascript, PHP and Perl.
4 *Mark-up* languages are somewhat different in nature, as they are designed to annotate text, to denote either structural elements ('this is an author name'), presentation ('print this in italic') or both. The first such languages were designed for formatting documents for printing; an example is LaTeX. More recent examples control the format and structure of web resources; examples are HTML and XML.

An innovation in the way software is provided has come from the *open source* movement. Commercial software is usually provided as a 'black box'; the user has no access to the program code, and therefore cannot modify the system at all, nor even know exactly how it works. With open source software, the user is given the full source code – the original programs – which they are free to modify. This makes it easy to customize software to meet local needs and preferences. It also allows users to collaborate in extending and improving the systems, correcting errors, etc.; this is held by open source enthusiasts to be the best way of getting good-quality software. There is an analogy with the open access and open data initiatives, discussed in Chapter 10. Most such software is free, leading it to be known as FOSS (Free and Open Source Software). Well known examples of open source systems are the Linux operating system and the Moodle virtual learning environment; library and information examples are the Greenstone and Koha library management systems, the DSpace repository system, the Alfresco document and records management system, and GIS packages such as QGIS and GRASS. For an overview of open source systems generally, see Deek and HcHugh (2008), for an overview of library and information applications, see Hale and Hughes (2012) and Poulter (2010), and for examples of the use of specific systems, see Donnelly (2010), Biswas and Paul (2010) and Keast (2011).

Another way of getting access to the workings of software systems, though this time without having access to the source code, is via an Application Programming Interface (API). This is a facility provided by many websites, allowing external users to make use of the site's facilities in specified ways. Ince (2011) gives examples, such as a website dealing with French food and recipes which uses the Amazon API to display French recipe books and allow their purchase from Amazon without leaving the food site. A widely used API is that of Google Maps, which can be used to create location guides on one's own website. APIs can also be used to create mashups by taking data from various sources without the need for programming.

Interacting with computers

The ways in which people interact with computers, generally described under the headings of *human-computer interaction* (HCI) or of *usability*, is an area of great and increasing importance to the information sciences, as our opening quotation from Jonathan Grudin exemplifies; indeed it has sometimes been held to be the most significant area of overlap between information technology and information science. It is particularly concerned with the design and evaluation of interfaces, and has contributed greatly to some of the newer design methods mentioned below. Its methods range from surveys and interviews to the use of sophisticated instruments to track eye and hand movements as an interface is used. For overviews of these areas, see Chowdhury and Chowdhury (2011), Rogers, Sharp and Preece (2011), Grudin (2011), Dix et al. (2004), and Shneiderman et al. (2009).

A related area is that of *information visualization*. Whereas HCI responds to the opportunities for different forms of interaction offered by present-day computers, visualization responds to the challenge of the great amounts of digital data which can be obtained. This involves the use of software to help in the comprehension of large volumes of data, often by creating visual or diagrammatic representations, and by extracting and displaying crucial facts and figures from the mass. For an overview and numerous examples, see Steele and Iliinsky (2010), and for briefer exemplification, see Davis and Shaw (2011, Chapter 8). There is a link between visualization and the broader area of *data analytics* or *data mining* – extracting interesting information from large data compilations – which will be mentioned in Chapter 12.

Information systems, analysis, architecture and design

The way in which IT systems are modelled and designed has changed greatly over the past decades. This is relevant to information science practitioners for two reasons: because it influences the kind of information systems which are available; and because some of the newer methods allow much more involvement from those who are not technical experts, including many information practitioners. For an overview of recent trends and issues see Galliers and Corrie (2011).

There is dichotomy in the general methods by which IT systems are envisaged and designed, between approaches which we may categorize, at the risk of oversimplification, as 'hard' and 'soft'. Hard approaches, including most systems analysis and design and software engineering, take a positivist viewpoint; the aspects of the world being considered are real and unambiguous, and amenable to formal data modelling, metrics, fixed objectives, and structured processes of analysis. Softer approaches, including soft systems methodologies, take a more interpretivist viewpoint: there are different perspectives of the aspects of the world being studied, and all should be taken into account, through a toolbox of methods including subjective conceptual models and an emphasis on understanding the problems qualitatively.

A classic overview and comparison of methods, with particular emphasis on soft systems approaches, is given by Checkland and Holwell (1998). Peter Checkland, a British management scientist, is the originator of the soft systems approach, and still explains it more clearly than later authors: for a recent and readable account of this approach, see Checkland and Poulter (2006). There are many texts dealing with the 'harder' approaches; helpful examples are Kendall and Kendall (2010) and Britton and Doake (2005).

New trends in system design, influenced by HCI concepts and methods, and by soft systems approaches, are variously termed 'user-centred', 'participatory', 'persuasive' and 'interactive'; see, for example, Rogers, Sharp and Preece (2011), Hasle (2011) and Shneiderman et al. (2009).They emphasize an incremental, interactive approach, changing the system to suit the needs of the user, rather than forcing the user to adapt to the 'best' system. Information practitioners are often well positioned to be involved in this sort of design process.

A similar aspect of use of IT in which information practitioners are often involved is that of *information architecture* (IA). The term has a number of shades of meaning, but it is usually understood as the quest for the best way of organizing and structuring complex information spaces, typically websites, so that their users can find their way around easily, and feel that the information environment is supportive of their needs. It draws from the insights of HCI, of user-centred design, and from the 'traditional' ideas of information design on the printed page. It also relies on the ideas of categorization, taxonomy, and faceted classification, discussed in Chapter 6 to organize the layout of information. The term was coined in the 1970s by Saul Wurman, originally a 'real' architect, who came to believe that the principles behind the design of buildings and physical spaces could be applicable to information spaces; he is, perhaps, better known for his concept of 'information anxiety'. Information architecture came into its own from the late 1990s, with the expansion of the web, and the need to bring order and structure to web-based information. Dillon (2002) gives an interesting brief account of issues in the development of the subject, and Haller (2011) gives a summary of its current status. The fullest introduction, emphasizing the application of IA to large websites, is Morville and Rosenfield (2006); Morville and Calender (2010) gives copious examples for search and navigation interfaces. See Davies (2011) for an account of the use of IA to counter information overload, and Burford (2011) for a study of the practice of IA in several large organizations. For the basic principles of information design in any context, and on which IA draws, see Orna (2005), and for an interesting historical perspective, see the analysis of Paul Otlet's ideas for interfaces to information by van den Heuvel and Rayward (2011). Although IA has no agreed general theory, some pragmatic principles have emerged: Brown's (2010) 'eight principles of information architecture', shown in slightly modified form in the box on the next page, are a good example.

Having dealt with a number of general issues, we can now turn to look at some specific applications.

Applications

We will briefly review some applications of information technologies of particular importance to the information sciences. We will do this by considering their impact on aspects of the communication chain – creation, dissemination, sharing, organization and retrieval, and preservation of information – bearing in mind that the impact will include digital documents that may be printed, and paper documents that may be digitized.

Creation

The most familiar form of technology for information creation comes in the well known form of office software: word processors, spreadsheets and presentation

Principles of information architecture (adapted from Brown, 2010)

Objects: treat content as a living thing, with a lifecycle, behaviours and attributes; recognize different types of content and treat them differently.

Choices: offer meaningful choices to users, keeping the range of choices available focused on a particular task. A greater number of options can make it more difficult for people to reach a decision: people think they like having many options, but they don't (the 'paradox of choice').

Disclosure: show only enough information to help people understand what kinds of information they will find if they dig deeper. Comes from the general design principle of 'progressive disclosure': people cannot use information they are not yet interested in, or do not understand.

Exemplars: describe the contents of categories by showing examples; it's the simplest and most effective form of explanation.

Front and side doors: assume a majority of users will come to any page or piece of information other than through the home page and prescribed navigation routes; typically they come via search engine. All pages should tell the visitor where they are, and what else is available; the home page should focus on orienting new users.

Multiple classification: offer several different classifications for browsing content; allow for the users' different mental models, even for quite restricted sets of information.

Focused navigation: don't mix apples and oranges in a navigation scheme; provide access by different mechanism, for example, topic, timeliness, services; use facet analysis principles.

Growth: assume the content you have today is a small fraction of what you will have in the future; allow for growth by, for example, having a few main categories and making it easy to create sub-categories.

software. To this, we might add software packages for editing images, video and audio files created by digital devices. These are used in the creation of most 'born digital' documents of a traditional kind.

There is then a very considerable body of information in the form of, largely ephemeral, digital communications: e-mails, micro-blogging messages, social media updates, and so on.

In a less familiar context, much digital data is gathered automatically by a wide variety of monitoring instruments: from CCTV cameras to satellite imaging, and from medical diagnostic data to retailers' sales data. These create the large data sets which are the raw material for the e-research to be discussed in Chapter 10.

Dissemination

IT has affected the dissemination aspect of the communication chain in a number of ways. Publishing has been affected in two main ways. Conventional printed documents are produced by digital printing, with the advantages noted above. And much publishing is now wholly digital, in the form of web-based publicly available materials, and in-house material published on intranets; both of these relying on what used to be termed *desktop publishing* facilities, essentially web design and content management software.

Newer forms of dissemination, most obviously blogs, rely on combining easy-to-use web page creation software, together with communication links. Mashups, referred to above, which link and integrate information from disparate sources by use of APIs and scripting languages, are a similar new form of dissemination; for an overview of mashups with library and information examples, see Engard (2009).

Another form of dissemination promoted by IT systems is digitization of physical items: books, newspapers and other texts, and also images, museum objects, etc. This makes large collections of older material, as well as individual rare and valuable items, much more widely available than their physical instantiation would permit. This process involves hardware (scanners) and software (image manipulation and content management); see Zhang and Gourley (2008) and Terras (2010) for overviews.

Finally, we should mention *machine translation*, as a means of disseminating material in languages unfamiliar to potential users, much more widely than would be possible with human translators. After many years of unfulfilled promise, this has now reached a stage where a useful, if not perfect, translation between many language pairs is available free, or at low cost, in part due to availability of systems such as Babelfish and Google Translate. For a detailed overview of machine translation see Wilks (2009), and for examples of library and information applications see Spellman (2011) and Smith (2006).

Sharing

The technologies which support information sharing come, for the most part, in two broad categories: one, relatively long-established and relatively formal, we can categorize as *groupware*; the other, its opposite in both respects, as *social media*.

Groupware, also referred to as *collaborative software* and as *computer-supported collaborative work* (CSCW) has its origins in experimental systems developed in the 1960s, but only achieved wide adoption in the 1990s. Its core is usually taken to be systems which allow collaborative work on documents by participants remote from each other, supported by a variety of tools for communication and conferencing and virtual meetings. In commercial environments, this is most often instantiated in packages such as Lotus Notes and Microsoft Sharepoint, but a wide variety of smaller systems of all kinds are used for collaborative working.

Social media originated with the world wide web in the 1990s, on which it depends, and became seemingly ubiquitous in the new millennium. Intended primarily for informal social interactions, through messaging and exchange of media such as photographs and videos, it has begun to be adopted for business and professional activities; for early analyses of its transformative effects, see Warr (2008) and Qualman (2009). It is exemplified by Facebook, primarily a social tool, though increasingly used for marketing and promotion, and by LinkedIn, a professional networking tool.

Spanning the boundaries of these two categories are *wikis* and *media sharing* sites. Wikis are web-based systems that allow their content to be extended and modified by their users, often with little or no central control. They are now used in many contexts: the best known is the Wikipedia public encyclopedia, but the approach is used for information sharing in many diverse contexts.

Media sharing sites on the web allow users to maintain collections of their own material while simultaneously making it available for others to access. Well known examples are Flickr for photographs and LibraryThing for books.

Finally, and at a more sophisticated level of sharing, we might mention *expert systems*, an aspect of artificial intelligence. These are software systems which aim to incorporate the knowledge of a community of expert people in a very limited area, and which then manipulate this knowledge through rules and other processes in order to act as an expert person would; making diagnoses, recommending a course of action, etc. Under development since the 1970s, and steadily increasing their use for very specific tasks, they have not really fulfilled their once much-hyped potential as yet. Human knowledge, especially of the commonsense variety, has proved surprisingly resistant to being computerized.

Organization and retrieval

These are the IT applications which have usually been thought of as at the heart of information science. Nonetheless, we will treat them, as with other applications, in outline only, as there are many excellent detailed texts available. They have a close relationship with the tools and concepts of information organization discussed in the last chapter, and with issues of information management discussed in Chapter 12; here we will focus on the technical issues. This is not to ignore the fact that studies of information retrieval, in particular, have generated interesting theoretical issues; see the discussions of paradigms in Chapter 3, an account of 17 theoretical constructs of information retrieval (Jansen and Rieh, 2010), and debates on the relevance of theories of knowledge to information retrieval, indexing and browsing (Hjørland, 2011a, 2011b).

We will discuss these applications under six headings, accepting that there is overlap between them: reference handling systems, databases, information retrieval systems, digital libraries and repositories, document management systems and the semantic web.

Reference handling software is the simplest form of these systems. It allows for collections of bibliographic or website references to be created and maintained, with limited indexing, and to be output in different citation formats. Examples are RefWorks and Endnote; Zotero, Mendeley and citeulike are web-based media sharing equivalents. See Kern and Hensley (2011) for a review of their features.

Databases is a rather over-used and general term for any collection of digital information. Here we will use it in a more restricted way, to mean systems which handle structured data, typically in the form of numbers and short pieces of text: the *database management system* (DBMS) or, in an even more restricted sense, the *relational database management system* (RDBMS).

The phrase 'relational database' strictly refers to a system designed with rigorous formal rules; but the term has been misused over the years by software companies who have applied it to systems which do not meet the full criteria. The key to RDBMSs is *normalization*. This is a process of grouping data elements into clearly defined structures without redundancy, according to a data model, removing repeating data elements so that each piece of data is stored only once, and making the model depend on the relations between the data elements, rather than any particular application of the data. The data is then represented as two-dimensional tables, where each table represents a kind of thing, each row one instance of that thing, and each column one attribute of that instance of the thing. So for example, we could have a table of BOOKS, with one column for AUTHOR, one column for TITLE, and one column for DATE OF PUBLICATION. Each row represents one book, e.g.

AUTHOR	TITLE	DATE PUBLISHED
Charles Dickens	A Christmas Carol	1843
Arthur Conan Doyle	A Study in Scarlet	1887

In a library setting we might make more tables, to deal with individual copies of the books and their location, borrower names and addresses, and so on. The important point is to remove multiple storing of the same information, and to have a logical structure so that only the same sort of information appears in the same tables; we would not, for example, want to have details of borrowers in the BOOKS table, as they would have to be repeated for each book they borrowed.

Database design can get complicated, as it has a good deal of associated theory, largely developed by the British computer scientists Edgar Codd and Chris Date, while they were working for IBM in the USA in the 1960s and 70s; however, the basic ideas are quite simple. The process of modelling for databases of this kind is termed *entity-relationship* (ER) modelling, and is the basis of the RDA metadata scheme discussed in Chapter 6, on information organization. Databases of this kind use a structured query language to search the tables; the best known is SQL (Structured Query Language) used in many database systems.

There are many books and articles dealing with DBMS, at a variety of levels of rigour and detail. Connolly and Begg (2010) is thorough and relatively accessible; for detailed treatments by one of the pioneers of the topic, see Date (2003; 2011). Davis and Shaw (2011, Chapter 7) give brief and clear examples.

Information retrieval (IR) systems are usually understood to be those which handle less structured data than the facts and figures dealt with by DBMS. There may be some structure present: for example, the field structure of a typical bibliographic database, with fields for author name, title of article, source name, publication details, subject indexing etc. Or their may be little or no structure to the records in the system, as with web search engines such as Google, or the 'enterprise search' systems, such as Autonomy, which aim to deal with large volumes of diverse materials, including reports, emails, and so on.

IR systems are usually analysed in terms of system components; however, it seems that each writer has their own preferred set of components. A recent, and relatively complex, example is shown in Figure 7.3.

We will discuss IR systems in a rather simple way, considering just four main components or sub-systems: input; indexing; search; and interface. Fuller coverage is given by Chowdhury (2010) and Chu (2010) from a library and information viewpoint, and by Baeza-Yates and Ribeiro-Neto (2010) in more technical detail. Other perspectives are given in the multi-authored books edited by Melucci and Baeza-Yates (2011), Ruthven and Kelly (2011), Foster and Rafferty (2011) and Göker and Davies (2009).

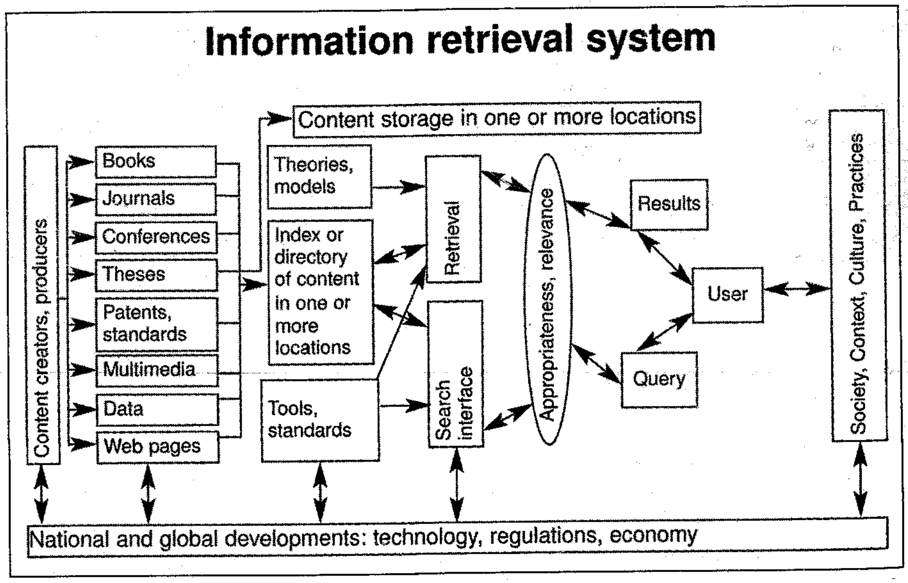

Information retrieval system
Content storage in one or more locations
Content creators, producers
Books
Journals
Conferences
Theses
Patents, standards
Multimedia
Data
Web pages
Theories, models
Index or directory of content in one or more locations
Tools, standards
Retrieval
Search interface
Appropriateness, relevance
Results
User
Query
Society, Context, Culture, Practices
National and global developments: technology, regulations, economy

Figure 7.3 *Components of an information retrieval system (reproduced from Chowdhury, 2010)*

The input sub-system, depending on the purpose of the overall system, may have to include materials of a widely different kind, as in the diagram in Figure 7.3, or may handle a more restricted range. What is entered may be a whole document, for a full-text retrieval system, or a document surrogate, a shorter summary record such as a bibliographic reference. Virtually all IR systems use an 'inverted file' structure; a file of index terms, created by the indexing sub-system, with pointers from each term to the documents to which it applies, themselves held in another file in sequential order. This allows rapid look-up of relevant documents, without having to scan the whole file.

The index file is created by the *indexing* component. Indexing may be 'full-text', with every word in the document indexed, or more selective; if the latter it may be automatic, with terms chosen statistically according to their frequency of occurrence, or it may be done by an indexer. A controlled vocabulary may or may not be used, as well as, or in addition to, the terms in the original document and/or added free terms. The process of indexing was discussed in the chapter on information organization.

The heart of an IR system is its *search* component, which does the work of retrieval. Search is carried out in different ways, reflected in the 'theories/models' component in the diagram in Figure 7.3; in essence it describes the nature of algorithms used in the search process. For details, see the recommended texts and the chapters by Rasmussen (2011) and Hiemstra (2009). We will mention three main classes of models:

1 *Boolean*: this works by manipulating sets of documents indexed with particular terms, using the Boolean operators AND, OR, NOT. It is clear and straightforward: a search for '(CATS OR DOGS) NOT HORSES' produces all those documents indexed with the term CATS or the term DOGS provided they are not indexed with the term HORSES. A drawbacks is the binary all-or-nothing metric. There is no way to say that one document is all about cats, while another barely mentions them in passing; documents are indexed CATS or not. Also, users often find it difficult to express their information need in the exact way required by a Boolean search.
2 *Vector*: this approach calculates the similarity between the query and each document according to the terms in common. It is more subtle than Boolean, as it allows for different degrees of matching, rather than just yes/no, and it allows results to be ranked in order of similarity to the query.
3 *Probabilistic*: this approach includes a number of methods which attempt to calculate the probability that a document will be relevant to a query, on the basis of statistical analysis of the occurrence of terms in the query and

in the set of documents. Some of the methods use 'relevance feedback', allowing the user to assess retrieved documents as relevant or not, amending probabilities according to these judgements, and searching again; this process continues until the user is satisfied with the result.

Sometimes the Boolean model is described as 'exact' retrieval, while the other two models are denoted as 'best match'; the latter will always return some result as the closest to the query, whereas the former will return exactly what was specified, which may be nothing.

Other retrieval algorithms use fuzzy set theory and Bayesian logic (Baeza-Yates and Ribeiro-Neto, 2010); there have even been studies on analogies between the formalisms of retrieval and quantum mechanics, with potential application in the quantum computers mentioned below (van Rijsbegen, 2004; Melucci and van Rijsbergen, 2011). Some retrieval systems use information other than apparent relevance to rank the output; an obvious alternative is to present newer material first. The Google search engine's PageRank algorithm orders items according to how many other web pages link to them, a surrogate measure of the interest value of the site; and one which has spawned a small industry of 'search engine optimization' to promote sites in these rankings.

Many writings on searching assume that the user has a well defined query to put to the system, but in fact many searches involve an element of *browsing*. There are numerous purposes for browsing, including:

- finding information in a context where browsing is the only feasible method
- finding information on topics which are not clearly defined, or which are hard to specify exactly; i.e. where the information need is broad and poorly specified
- getting an overview or sample of the information in a collection
- finding items which are similar to, or dissimilar from, those which one has identified
- finding one's bearing in a subject of which one knows little
- selecting the 'right' information from a large collection of 'relevant' material
- looking for inspiration, new ideas, or just something interesting; i.e. allowing for serendipity.

Browsing has been included as a component of a number of models of information-seeking behaviour, as we will discuss in Chapter 9, although usually without clear definition. It is variously categorized as *directed*, *semi-directed* and *undirected*, as *systematic*, *exploratory* and *casual*, or as *purposive*, *semi-purposive* and *capricious*, and has also been described by terms such as *undirected viewing*, *active scanning* and *passive attention*. Some IR systems make specific provision for this kind of 'informal search', but most do not cater for it well (Bawden, 2011).

The *interface* component allows the user to put their query to the system, and to receive the results. Interfaces have changed greatly over the years, although the operations which they evoke behind the scenes have not. There have been, and still are, a variety of interface styles; see the recommended texts, Morville and Callender (2010) and Wilson (2011) for details and examples. We will mention three important general styles, which have developed to match the typical computer interfaces of the time.

The first was the *command line* interface, stemming from the 1960s and developed for the first generation of online bibliographic search systems. At least one of these, for the Dialog host system, remained in use almost unchanged over nearly 50 years. These require the user to know a limited set of commands: typically, to search for documents indexed with a term, browse the index, combine sets with Boolean operators, and display results. They also allow for searching for phrases, truncated terms beginning or ending in a certain way, terms within a specified proximity of one another, terms in particular fields of the record, and so on. For instance, in the classic Dialog system, the command:

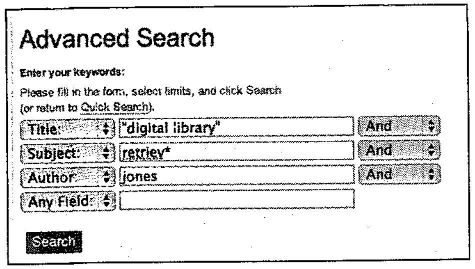
```
S digital(w)library/ti AND (online OR retrieval OR search*) AND au=jones ?
```

will retrieve documents with the phrase 'digital libraries' in the title, and any of the terms 'online', 'retrieval' or anything beginning 'search', for example search, searches, searching, and so on, anywhere in the subject parts of the record, and an author with the name Jones.

This kind of interface gives the user complete control over a Boolean search, and allows very precise specification of detailed searches. However, it not easy to use, and the user must take some time learning the system commands. Although these types of interfaces are still available, their functions have largely been absorbed within other styles of interface.

At the opposite extreme is the simplicity of the 'search box', pioneered by Google; a single box, into which the user types whatever they wish. This may be a Boolean statement, if the system works in this way, but will more usually be just a few terms or phrases; most searches on Google are single words. This simplicity is so attractive to most users, particularly if combined with a search function which gives ranked output, that many IR systems have adopted it, accepting that it loses the control and precision possible with the command line.

An intermediate stage, often provided as an 'advanced search' function, attempts to combine the simplicity of the search box with the power of the command line, by providing a matrix of rows and columns for the user to fill in, assisted by prompts and drop-down menus, giving a limited choice of Boolean operators and field specification. The example below, Figure 7.4, from our university's OPAC, shows a search for 'digital library' as a phrase in the title, the truncated term 'retriev' in subject terms, and an author name Jones.



Advanced Search

Enter your keywords:

Please fill in the form, select limits, and click Search (or return to Quick Search).

Title: "digital library" And

Subject: retriev* And

Author: jones And

Any Field:

Figure 7.4 *Advanced search function*

All these forms of interface, and variants of them, are to be found in IR systems; system designers seek to strike a balance between simplicity and precision of search specification; most users prefer simplicity most of the time.

So far, we have discussed IR systems in terms of text retrieval, by contrast with the facts and figures searching of DBMS. Although this is certainly the most widely encountered form of IT system, there are several other forms for different types of material, for example:

- citation searching, finding which documents cite a starting point document, to find newer material on that subject (Jacsó, 2004)
- image searching, using algorithms to search for features of still and moving images, rather than relying on text indexing (Enser, 2008a; Enser, 2008b; Town and Harrison, 2010)
- sound searching, to find items with names that sound similar; for example to identify medicines when all that is known is the spoken name (Robinson, 2010)
- music retrieval, with algorithms searching for features of pieces of music, rather than relying on index terms (Downie, 2003; Inskip, 2011)
- cross-language information retrieval (CLIR) (Kishida, 2005)
- chemical structure and substructure searching, retrieving records for chemical substances and reactions (Willett, 2008).

There have been many studies of the ways in which people use retrieval systems; a detailed aspect of the information behaviour which we will discuss in Chapter 9. Initially focused on system aspects, these have recently turned to examining individual preferences and styles; as examples, see Ford, Miller and Moss (2005), Chen, Magoulas, and Dimakopoulos (2005) and Vilar and Žumer (2008). Heinström (2010) has used such studies to define searching 'styles', such as the conscientious searcher, the worried searcher, and the laid-back searcher.

Studies of the use of IR systems have identified a small number of search strategies and tactics, general approaches to retrieval, used frequently; these can be put to best effect with those interfaces which give full control over Boolean searching in structured records with subject indexing. They were first described by Marcia Bates (1979), whom we met in Chapter 4; for recent comparisons, see Papaioannou et al. (2010) and Xie (2010), and for similar tactics with internet search engines see Smith (2012). A recent development has been the trend to 'social discovery' or 'social search', with search becoming a collaborative activity through means such as social bookmarking and tagging (Shneiderman, 2011; McDonnell and Shiri, 2011).

Some examples of identifiable strategies are shown in the box below: in a practical setting they will mainly be used informally and in combination, but it may be helpful to adopt them if no useful material, or too much, is being found.

Examples of search strategies for information retrieval

Berrypicking: carry out the search using a small number of search terms; choose the most interesting two or three items; read these, and then search again modifying the search terms as necessary; repeat until enough items have been found (the topic of the search may have changed in the process).

Building blocks: divide the search into distinct topics, represent each topic by a set of synonyms linked by OR and then link all the topics with AND. Begin with the topic likely to have the smallest number of documents in the collection, and stop when the number of retrieved documents is small enough to scan.

Pearl growing: start with a known relevant document, the 'seed', and look it up in the collection. Examine the index terms, and use these as starting points for a search. Examine the retrieved documents, choose relevant one, and repeat the process. Continue until no new relevant documents are found.

Successive fractions: do an initial search, using just one term, or a few terms linked with AND. Examine some of the documents found, and find terms occurring only in those which are relevant or those which are not relevant. Repeat the search, adding 'relevant' terms with AND and 'not relevant' terms with NOT. Continue until the retrieved set has been reduced to a small enough number to be scanned.

Quicksearch: divide the search into distinct topics, represent each with one index term, and search linking the terms with AND. Any relevant documents found can then be used as the 'seed' for pearl growing.

The evaluation of IR systems has been a major preoccupation of information science over many years, and has spawned its own traditions, methods and metrics; see, for example, Robertson (2008). The most widely used metrics for information retrieval are recall and precision, measuring respectively the success of a system in finding all the relevant material that there is to be found, and in returning only relevant material. They are formally defined as follows:

For a search which retrieves N documents, of which Nr are judged relevant, and there are M relevant documents in the collection:

Recall = Nr / M
Precision = Nr / N

See Robertson (1969) for a classic paper introducing this kind of metrics and Egghe (2008) and Järvelin (2011) for more recent analyses. See Chapter 14 for methods of evaluating IR systems, and Chapter 4 for a discussed of the complex question of relevance, on which all these metrics depend.

Digital libraries and repositories

Strictly speaking, a digital library would be a library whose collections and services were wholly digital, having no physical location, and being staffed by virtual librarians; perhaps Second Life avatars. No libraries have yet achieved this state, and few libraries have only digital material. By common usage, a digital library is a library offering a significant proportion of digital material and services, along with physical components; these have also been termed electronic, hybrid and virtual libraries. For explanations and analyses of the complex nature of digital libraries, and the way in which these have developed over time, see Rowlands and Bawden (1999), Bawden and Rowlands (1999), Arms (2000), Seadle and Greifeneder (2007). Candela et al. (2007) give a detailed set of definitions and models for these systems and their components.

All digital libraries provide the usual range of library services, which will be outlined in Chapter 12, but with particular differences to the way the collection is defined and maintained. The diagram in Figure 7.5 indicates the major components of a digital library, in particular showing how such a library 'extends' its collections into web resources and in the collections of other digital libraries. The diagram shows the most common arrangement, by which each form of material has its own search interface/s; there may, for example, be different interfaces for groups of databases and collections of e-journals, as well as for special collections and materials, including printed materials. The alternative, a single interface for all forms of material, is desirable, but difficult to arrange, if it is to avoid an overly simple 'lowest common denominator' approach.

Digital libraries grew from initial development in library automation (see Tedd (2007) for the early history of this in the UK), given impetus by the development of the web, and increasing availability of digital forms of traditional resources, particularly e-journals. Overviews and examples are given by Arms (2000), Chowdhury and Chowdhury (2002), Andrews and Law (2004), Lesk (2005), Bearman (2007) and Chowdhury et al. (2008). There are particular concerns about *interoperability*, the ability to search across different collections and different libraries, requiring common standards and metadata, and *link resolver* facilities, to enable a user to switch between, for example, a record which they have found in an abstracting service and the equivalent full-text article in an e-journal collection.

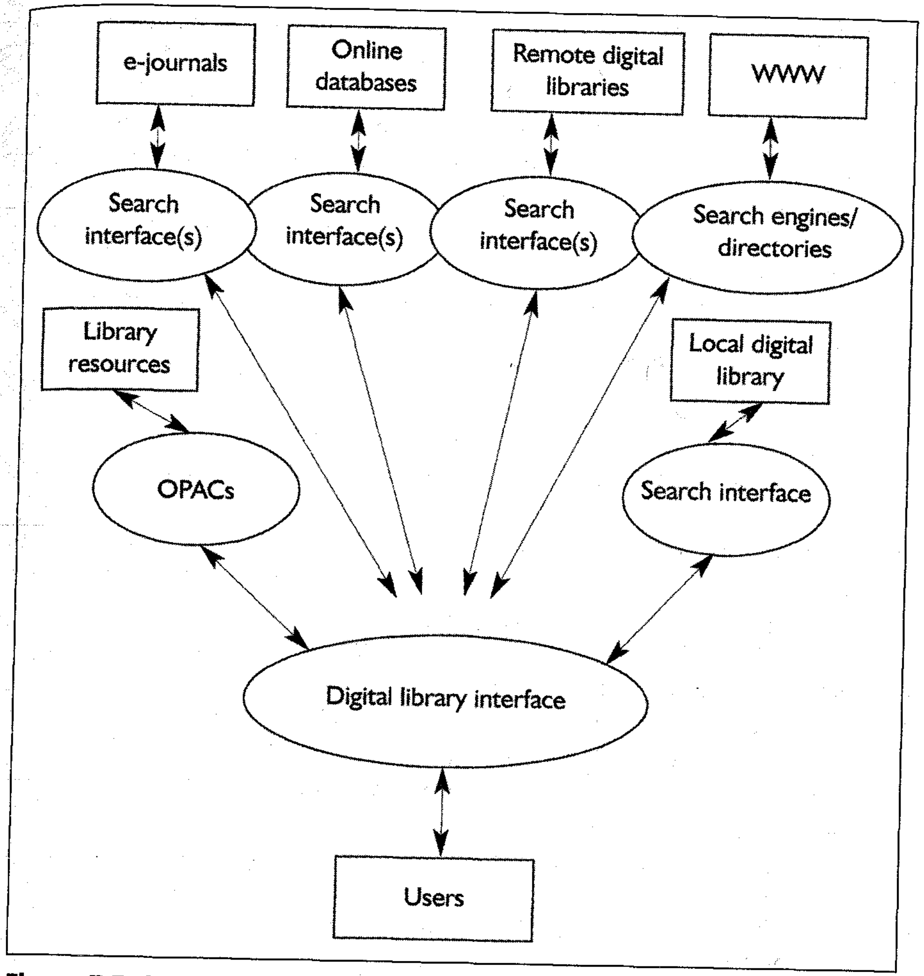


Figure 7.5 *Components of a digital library (reproduced from Chowdhury and Chowdhury, 2002)*

Closely associated with digital libraries are *library management systems*, which can be used with both digital and printed resources to automate all library management and service functions, and which are generally structured around an online catalogue (OPAC). Current examples include: open source systems such as Koha and Greenstone: proprietary systems such as Talis, Ex Libris (Aleph and Voyager) and SirsiDynix: and the media-sharing web-based LibraryThing, designed for personal book collections but now being increasingly used in 'proper' libraries.

Repositories, which we will discuss in Chapter 10 in the context of open access, are systems for maintaining collections of documents created within an organization and making them publicly available. They are similar in concept to digital libraries, their uniqueness lying in the form of material, which means that collection building is a rather simple process, and that indexing and classification can be relatively simple. They typically use software systems specifically designed for repositories, such as DSpace and EPrints.

Electronic document and records management systems (EDRMS) handle the information flows within records centres and archives; as we will see in Chapter 12, a main feature of this environment is a clear document lifecycle, which these systems are designed to manage. Examples are HP TRIM, ECM Documentum, and the cloud-based Alfresco.

Finally, although it is not a system as such, but rather an aspiration, we should mention the *semantic web*. This refers to the idea that information on the web may be structured and encoded in such as way that its content and meaning are made explicit, so that they can be 'understood' by search engines and other software agents. The idea was first popularized by Tim Berners-Lee, the originator of the web itself, at the start of the new century (see, for example, Berners-Lee, Hendler and Lassila, 2001). It is still very much a work in progress, relying on the developments in metadata, ontologies, taxonomies, tagging, etc., discussed in Chapter 6. The concept of *linked data*, formal means by which related items can be associated by structured metadata, is also fundamental. For progress reports on its development, and relevance to information provision, see Antoniou and van Harmelen (2008), Burke (2009), Dunsire and Willer (2011) and Miller (2011, Chapter 11).

Preservation

The final stage of the communication chain, archival preservation, is both helped and hindered by information technologies. On the one hand, having digital copies of printed materials is a good safeguard against loss through fire, flood, theft, etc. On the other hand, the preservation of digital materials is in itself a major problem, and one so new that the best methods are not yet fully understood. For one thing, whereas paper documents are well known to survive for hundreds of years if properly cared for, no digital storage format has been in existence for more than 50 years, so that long-term survival cannot be demonstrated. There is also the problem of the obsolescence of formats. As storage mechanisms pass from use – who now remembers the Betamax video tape or the 8-inch floppy disk? – the data on them must be migrated to newer formats. Or the old equipment must be archived along with its data. Not for nothing do some archivists argue that the best way to ensure the survival of digital data is to print it on paper and put it in a drawer. For an overview of these issues, see Harvey (2010; 2012).

Summary

Future prospects for information technology largely stem from the seemingly inexorable increases in processing power and storage capacity; as we shall see in Chapter 15, some commentators believe that this will lead, probably around the middle of the 21st century, to a 'singularity'; a point at which the power of computers becomes so great that humans will no longer be able to understand what they are doing. Even setting these aside, possible advances include: quantum computers, based on the 'qubit', a storage element able to be in several states at once, and offering greatly enhanced processing capabilities; biological computers, using strands of DNA as the processing elements; neural computers, designed to emulate the patterns of neurons within the human brain; and various flavours of artificial intelligence, so that the behaviour of the computer, in a particular context, would be indistinguishable from that of a person.

For the moment, we can conclude that, whether or not any of these ideas comes to fruition, information technology will continue to develop and continue to affect and change the communication chain of recorded information. Information science and information technologies are intimately linked, each continually influencing the other. Researchers and practitioners in the information sciences can be designers, as well as users, of technology systems, particularly in HCI and information architecture, as well as in the obviously relevant applications in the organization and retrieval of information.

- Computers are still designed according to the concepts put forward by von Neumann in the 1940s, though their power and speed continue to increase exponentially.
- Networking and pervasive computing have greatly altered, and continue to alter the IT environment.
- All aspects of the communication chain of recorded information are affected by developments in IT.
- Information practitioners have traditionally focused on IT applications such as information retrieval and digital libraries, but are well placed to contribute to other aspects of IT development.

Key readings

Darrel Ince, *The computer: a very short introduction*, Oxford: Oxford University Press, 2011.

[A clear and concise introduction to digital computers and networks.]

Gobinda Chowdhury, *Introduction to modern information retrieval* (3rd edn), London:

Facet Publishing, 2010.

[Gives a wide coverage of many of the topics of this chapter.]

Peter Morville and Louis Rosenfeld, *Information architecture for the world wide web: designing large-scale web sites*, Sebastopol CA: O'Reilly Media, 2006.

[Good introduction to all aspects of information architecture.]

Peter Morville and Jeffery Callender, *Search patterns*, Sebastopol CA: O'Reilly Media, 2010.

[Copious examples of search interfaces and features.]

References

Ally, M. and Needham, G. (eds) (2012) *M-libraries 3: Transforming libraries with mobile technology*, London: Facet Publishing.

Andrews, J. and Law, D. (eds) (2004) *Digital libraries: policy, planning and practice*, Aldershot: Ashgate.

Antoniou, G. and van Harmelen, F. (2008) *A semantic web primer* (2nd edn), Cambridge MA: MIT Press.

Arlitsch, K. (2011) The Espresso Book Machine: a change agent for libraries, *Library Hi Tech*, 29(1), 62–72.

Arms, W. Y. (2000) *Digital libraries*, Cambridge MA: MIT Press.

Baeza-Yates, R. and Ribeiro-Neto, B. (2010) *Modern information retrieval: the concepts and technology behind search* (2nd edn), Harlow: Addison-Wesley.

Bates, M. J. (1979) Information search tactics, *Journal of the American Society for Information Science*, 30(4), 205–14.

Bawden, D. (2011) Encountering on the road to Serendip? Browsing in new information environments, in Foster, A. and Rafferty, P. (eds), *Innovations in information retrieval*, London: Facet Publishing, 1–22.

Bawden, D. and Rowlands, I. (1999) Digital Libraries: assumptions and concepts, *Libri*, 49(4), 181–91.

Bearman, D. (2007) Digital libraries, *Annual Review of Information Science and Technology*, 41, 223–72.

Berners-Lee, T. (1999) *Weaving the web: the past present and future of the world wide web by its inventor*, London: Orion.

Berners-Lee, T., Hendler, J. and Lassila, O. (2001) The semantic web, *Scientific American*, May 2001 issue, 34–43.

Biswas, G., and Paul, D. (2010) An evaluative study on the open source digital library softwares for institutional repository: special reference to Dspace and Greenstone digital library, *International Journal of Library and Information Science*, 2(1), 1–10.

Britton, C. and Doake, J. (2005) *Software system development: a gentle introduction* (4th edn), Columbus OH: McGraw Hill.

Brown, D. (2010) Eight principles of information architecture, *Bulletin of the American Society for Information Science and Technology*, 36(6), 30–4.

Burford, S. (2011) Complexity and the practice of web information architecture, *Journal of the American Society for Information Science and Technology*, 62(10), 2024–37.

Burke, M. (2009) The semantic web and the digital library, *Aslib Proceedings*, 61(3), 316–22.

Candela, L. , Castelli, D., Pagano, P., Thanos, C., Ioannidis, G., Koutrika, G., Ross, S., Schek, H., and Schuldt, H. (2007) Setting the foundations of digital libraries: the DELOS manifesto, *D-Lib Magazine*, 13(3/4), [online] available at: http://www.dlib.org/dlib/march07/castelli/03castelli.html.

Checkland, P. and Holwell, S. (1998) *Information, systems and information systems: making sense of the field*, Chichester: Wiley.

Checkland, P. and Poulter, J. (2006) *Learning for action: a short definitive account of soft systems methodology and its use, for practitioners, teachers and students*, Chichester: Wiley.

Chen, S. Y., Magoulas, G. D. and Dimakopoulos, D. (2005) A flexible interface design for Web directories to accommodate different cognitive styles, *Journal of the American Society for Information Science and Technology*, 56(1), 70–83.

Chowdhury, G. G. (2010) *Introduction to modern information retrieval* (3rd edn), London: Facet Publishing.

Chowdhury, G. G., Burton, P. F., McMenemy, D. and Poulter, A. (2008) *Librarianship: an introduction*, London: Facet Publishing.

Chowdhury, G. G. and Chowdhury, S. (2002) *Introduction to digital libraries*, London: Facet Publishing.

Chowdhury, G. G. and Chowdhury, S. (2011) *Information users and usability in the digital age*, London: Facet Publishing.

Chu, H. (2010) *Information representation and retrieval in the digital age* (2nd edn), Medford NJ: Information Today.

Connolly, T. M. and Begg, C. E. (2010) *Database systems: a practical approach to design, implementation and management* (5th edn), Boston MA: Addison-Wesley.

Cope, B and Phillips, A. (eds) (2006) *The future of the book in the digital age*, Oxford: Chandos.

Date, C. J. (2003) *An introduction to database systems* (8th edn), Boston MA: Addison-Wesley.

Date, C. J. (2011) *SQL and relational theory* (2nd edn), Sebastapol CA: O'Reilly.

Davies, N. (2011) Information overload, reloaded, *Bulletin of the American Society for Information Science and Technology*, 37(5), 45–9.

Davis, C. H. and Shaw, D. (eds) (2011) *Introduction to information science and technology*, Medford NJ: Information Today.

Deek, F. P. and McHugh, J. A. (2008) *Open source: technology and policy*, New York NY: Cambridge University Press.

Derfler, F. J. and Freed, L. (2004) *How networks work* (7th edn), Indianapolis IN: Que.

Dillon, A. (2002) Information architecture in 'JASIST'; just where did we come from, *Journal of the American Society for Information Science and Technology*, 53(10), 821–3.

Dix, A., Finlay, J., Abowd, G. A. and Beale, R. (2004) *Human-computer interaction* (3rd edn), Harlow: Pearson.

Donnelly, F. P. (2010) Evaluating open source GIS for libraries, *Library Hi Tech*, 28(1), 131–51.

Downie, J. S. (2003) Music information retrieval, *Annual Review of Information Science and Technology*, 37, 295–340.

Dunsire, G. and Willer, M. (2011) Standard library metadata models and structures for the Semantic Web, *Library Hi Tech News*, 28(3), 1–12.

Egghe, L. (2008) The measures precision, recall, fallout and miss as a function of the number of retrieved documents and their mutual interrelations, *Information Processing and Management*, 44(2), 856–76.

Ekart, D. F. (2011) Tech tips for every librarian: Codify your collection, *Computers in Libraries*, 31(3), 38–9.

Engard, N. C. (ed.) (2009) *Library mashups: exploring new ways to deliver library data*, London: Facet Publishing.

Enser, P. (2008a) The evolution of visual information retrieval, *Journal of Information Science*, 34(4), 531–46.

Enser, P. (2008b) Visual information retrieval, *Annual Review of Information Science and Technology*, 42, 3–42.

Ford, N., Miller, D. and Moss, N. (2005) Web search strategies and human individual differences: a combined analysis, *Journal of the American Society for Information Science and Technology*, 56(7), 757–64.

Foster, A. and Rafferty, P. (eds) (2011) *Innovations in information retrieval*, London: Facet Publishing.

Galliers, R. D. and Corrie, W. L. (ed.) (2011) *The Oxford Handbook of Management Information Systems*, Oxford: Oxford University Press.

Göker, A. and Davies, D. (eds) (2009) *Information retrieval: searching in the 21st century*, Chichester: Wiley.

Gralla, P. (2006) *How the Internet works* (8th edn), Indianapolis IN: Que.

Grudin, J. (2011) Human-computer interaction, *Annual Review of Information Science and Technology*, 45, 369–430.

Hale, M. and Hughes, M. (2012) *Open source systems in libraries*, London: Facet Publishing.

Haller, T. (2011) Is information architecture dead?, *Bulletin of the American Society for Information Science and Technology*, 38(1), 42–3.

Hally, M. (2005) *Electronic brains: stories from the dawn of the computer age*, London: Granta.

Harvey, R. (2010) *Digital curation: a how-to-do-it manual*, London: Facet Publishing.

Harvey, R. (2012) *Preserving digital materials* (2nd edn), Berlin: de Gruyter.

Hasle, P. (2011) Persuasive design: a different approach to information systems (and information), *Library Hi-Tech*, 29(4), 569–72.

Heinström, J. (2010) *From fear to flow: personality and information interaction*, Oxford: Chandos.

Hiemstra, D. (2009) Information retrieval models, in Göker, A. and Davies, D. (eds), *Information retrieval: searching in the 21st century*, Chichester: Wiley, 1–19.

Hjørland, B. (2011a) The importance of theories of knowledge: indexing and information retrieval as an example, *Journal of the American Society for Information Science and Technology*, 62(1), 72–7.

Hjørland, B. (2011b) The importance of theories of knowledge: browsing as an example, *Journal of the American Society for Information Science and Technology*, 62(3), 594–603.

Hoy, M. B. (2011) An introduction to QR codes: linking libraries and mobile patrons, *Medical Reference Services Quarterly*, 30(3), 295–300.

Ince, D. (2011) *The computer: a very short introduction*, Oxford: Oxford University Press.

Inskip, C. (2011) Music information retrieval research, in Foster, A. and Rafferty, P. (eds), *Innovations in information retrieval*, London: Facet Publishing, 69–84.

Jacsó, P. (2004) Citation searching, *Online Information Review*, 28(6), 454–60.

Jansen, B. J. and Rieh, S. Y. (2010) The seventeen theoretical constructs of information searching and information retrieval, *Journal of the American Society for Information Science and Technology*, 61(8), 1517–34.

Järvelin, K. (2011) Evaluation, in Ruthven, I. and Kelly, D. (eds) (2011) *Interactive information seeking, behaviour and retrieval*, London: Facet Publishing, 113–38.

Keast, D. (2011) A survey of Koha in Australian special libraries: open source brings new opportunities to the outback, OCLC *Systems & Services: International Digital Library Perspectives*, 27(1), 23–39.

Kendall, K. E. and Kendall, J .E. (2010) *Systems analysis and design* (8th edn), Boston: Pearson.

Kern, M. K. and Hensley, M. K. (2011) Citation management software: features and futures, *Reference and User Services Quarterly*, 50(3), 204–8.

Kishida, K. (2005) Technical issues of cross-language information retrieval: a review, *Information Processing and Management*, 41(3), 433–55.

Krajewski, M. (2011) *Paper machines: about cards and catalogs, 1548–1929*, Cambridge MA: MIT Press.

Lesk, M. (2005) *Understanding digital libraries* (2nd edn), Waltham MA: Morgan Kaufmann.

McDonnell, M. and Shiri, A. (2011) Social search: a taxonomy of, and a user-centred approach to, social web search, *Program*, 45(1), 6–28.

Melucci, M. and Baeza-Yates, R. (eds) (2011) *Advanced topics in information retrieval*, Berlin: Springer.

Melucci, M. and van Rijsbergen, K. (2011) Quantum mechanics and information retrieval, in Melucci, M. and Baeza-Yates, R. (eds), *Advanced topics in information retrieval*, Berlin: Springer, 125–55.

Miller, S. J. (2011) *Metadata for digital collections: a how-to-do-it manual*, London: Facet Publishing.

Morville, P. and Callender, J. (2010) *Search patterns*, Sebastopol CA: O'Reilly Media.

Morville, P. and Rosenfeld, L. (2006) *Information architecture for the world wide web: designing large-scale web sites*, Sebastopol CA: O'Reilly Media.

Moulaison, H. L. and Corrado, E. (eds) (2011) *Getting started with cloud computing*, London: Facet Publishing.

Negroponte, N. (1995) *Being digital*, London: Hodder and Stoughton.

Orna, E. (2005) *Making knowledge visible*, Aldershot: Gower.

Palmer, M. (2009) *Making the most of RFID in libraries*, London: Facet Publishing.

Papaioannou, D., Sutton, A., Carroll, C., Booth, A. and Wong, R. (2010) Literature searching for social science systematic reviews: consideration of a range of search techniques, *Health Libraries and Information Journal*, 27(2), 114–22.

Poulter, A. (2010) Open source in libraries: An introduction and overview, *Library Review*, 59(9), 655–61.

Qualman, E. (2009) *Socialnomics*, Hoboken NJ: Wiley.

Rasmussen, E. (2011) Access models, in Ruthven, I. and Kelly, D. (eds) (2011) *Interactive information seeking, behaviour and retrieval*, London: Facet Publishing, 95–111.

Robertson, S. E. (1969) The parametric description of retrieval tests, *Journal of Documentation*, 25(1), 1–27.

Robertson, S. E. (2008) On the history of evaluation in IR, *Journal of Information Science*, 34(4), 439–56.

Robinson, L. (2010) *Understanding healthcare information*, London: Facet Publishing.

Rogers, Y., Sharp, H. and Preece, J. (2011) *Interaction design: beyond human-computer interaction* (3rd edn), Chichester: Wiley.

Rowlands, I. and Bawden, D. (1999) Digital Libraries: a conceptual framework, *Libri*, 49(4), 192–202.

Ruthven, I. and Kelly, D. (eds) (2011) *Interactive information seeking, behaviour and retrieval*, London: Facet Publishing.

Seadle, M. and Greifeneder, E. (2007) Defining a digital library, *Library Hi-Tech*, 25(2), 169–73.

Shneiderman, B. (2003) *Leonardo's laptop: human needs and the new computing technologies*, Cambridge MA: MIT Press.

Shneiderman, B. (2011) Social discovery in an information abundant world: designing to create capacity and seek solutions, *Information Services and Use*, 31(1), 3–13.

Shneiderman, B., Plaisant, C., Cohen, M. and Jacobs, S. (2009) *Designing the user interface: strategies for effective human-computer interaction* (5th edn), Boston MA: Addison-Wesley.

Smith, A. G. (2012) Internet search tactics, *Online Information Review*, 36(1), 7–20.

Smith, D. A. (2006) Debabelizing libraries: Machine translation by and for digital collections, *D-Lib Magazine*, 12(3) [online] available at http://www.dlib.org/dlib/march06/smith/03smith.html.

Spellman, R. (2011) Developing best practices for machine translation using Google Translate and OCR terminal, *Journal of Interlibrary Loan, Document Delivery and Electronic Reserve*, 21(3), 141–47.

Steele, J. and Iliinsky, N. (2010) *Beautiful visualization*, Sebastapol CA: O'Reilly Media.

Tedd, L. A. (2007) Library management systems in the UK: 1960s–1980s, *Library History*, 23(4), 301–16.

Terras, M. M. (2010) The rise of digitization: an overview, in Rikowski, R. (ed.), *Digitisation perspectives*, Rotterdam: Sense Publishers.

Town, C. and Harrison, K. (2010) Large-scale grid computing for content-based image retrieval, *Aslib Proceedings*, 62(4–5), 438–46.

Twyman, M. (1998) *British Library guide to printing history*, London: British Library.

van den Heuvel, C. and Rayward, W. B. (2011) Facing interfaces: Paul Otlet's visualizations of data integration, *Journal of the American Society for Information Science and Technology*, 62(12), 2313–26.

van Rijsbergen, C. J. (2004) *The geometry of information retrieval*, Cambridge: Cambridge University Press.

Vilar, P. and Žumer, M. (2008) Perceptions and importance of user friendliness of IR systems according to users' individual characteristics and academic disciplines, *Journal of the American Society for Information Science*, 59(12), 1995–2007.

Walsh, A. (2011) Blurring the boundaries between our physical and electronic libraries: location-aware technologies, QR codes and RFID tags, *Electronic Library*, 29(4), 429–37.

Warr, W. A. (2008) Social software: fun and games or business tools?, *Journal of Information Science*, 34(4), 591–604.

Whitchurch, M. J. (2011) QR codes and library engagement, *Bulletin of American Society for Information Science and Technology*, 38(1), 14–17.

White, R. and Downs, T. E. (2007) *How computers work* (9th edn), Indianapolis IN: Que.

Wilks, Y. (2009) *Machine translation: its scope and limits*, Dordrecht: Springer.

Willett, P. (2008) From chemical documentation to chemoinformatics: 50 years of chemical information science, *Journal of Information Science*, 34(4), 477–99.

Wilson, M. (2011) Interfaces for information retrieval, in Ruthven, I. and Kelly, D. (eds), *Interactive information seeking, behaviour and retrieval*, London: Facet Publishing, 139–70.

Wisniewski, J. (2011) Mobile usability, *Bulletin of the American Society for Information Science and Technology*, 38(1), 30–2.

Xie, I. (2010) Information searching and search models, *Encyclopedia of Library and Information Sciences* (3rd edn), London: Taylor & Francis, 1:1, 2592–2604.

Zhang, A. B. and Gourley, B. (2008) *Creating digital collections: a practical guide*, Oxford: Chandos.

Zimerman, M. (2011) Radio frequency identification (RFID): Time to take another look, OCLC *Systems and Services: International Digital Library Perspectives*, 27(2), 146–54.

Zorkoczy, P. (1982) *Information technology: an introduction*, London: Pitman.

CHAPTER 8

Informetrics

To measure is to know.

If you can not measure it, you can not improve it.

William Thomson, Lord Kelvin

The only man I know who behaves sensibly is my tailor: he takes my measurements anew each time he sees me. The rest go on with their old measurements and expect me to fit them.

George Bernard Shaw

With an uninformed reading, and taking information to be a basic constituent of the universe, informetrics seems a very broad subject. A more informed reading might narrow informetrics to the study of all quantifiable aspects of information science. The reality is . . . more modest still.

Concepción Wilson (1999, 107)

Introduction

This chapter is about measurement; specifically measurement of the quantitative aspects of the creation, communication and use of information. As the quotations above remind us, measurement is vital, but only if the right things are measured and the measurements are meaningful and up to date.

In this chapter, we will first examine the nature of informetrics and some of its components – bibliometrics, webometrics, etc. – before looking at how the subject has developed since its origins in the 1920s. We will consider the very basic question of how much information there is before looking, in a largely qualitative way, at the main 'informetric laws'; Lotka, Bradford and Zipf, and their offspring. Finally, we will look at how informetrics techniques may be applied in information research and practice, giving just an overview with examples.

Informetrics is the study of the quantitative aspects of information resources and of the communication of information. The term was introduced in 1979 by