1 Introduction

With the development of the World Wide Web, cheap CD-ROMs and, generally, easy access to computing and communication technologies, it has become easy to publish collections of documents, so that a huge variety are now accessible to a wide audience. Some of this variety is shown in Table 1.1.

Table 1.1 Some kinds of collections of documents.

- Books – both libraries and bookstores
- Articles – newspapers and magazines
- Memos and e-mail messages in large organizations
- Abstracts – a wide range of bibliographic databases
- Legal cases and statutes in many jurisdictions
- Web sites – organized by search engines collections of databases
- Images
- Museum collection catalogues
- Music fragments
- Criminal information, both on-going police records and specific investigations, including mug shots and fingerprints
- Records/CDs
- Videos, movies
- TV news footage
- Software libraries
- Queries, problems, frequently asked questions for software products
- Submissions to public enquiries
- Telephone hotline calls for product contamination incidents

People sometimes use these collections to find particular documents according to various criteria, and sometimes want to find out general characteristics of the collection, either as some form of statistical content analysis or some type of visualization. Publishing the documents therefore includes provision of access to information retrieval, analysis and visualization technology of various kinds.

Provision of technology is easy and it makes life easier for the users of the document collections, but its utility is limited if it is used on the raw document collection. The problems are not deficiencies in the technology, but are inherent in the way the documents are created and searched. At present, nearly all documents are texts in natural language. Even collections of images are generally represented by catalogue entries and captions, which are searched instead of the image itself. Different documents are created by different people for different purposes, so the vocabulary used and the assumed context differ from document to document.

Further, the person searching the collection generally has a different purpose in mind than any of the document creators.

Of course, even though many people and organizations are publishing documents for the first time, the problem of managing and organizing large collections of documents is not new. Librarians have been managing public collections for hundreds of years, and commercial bibliographic databases have been available for decades. These collections have historically been managed by information science professionals of various kinds, including librarians, records managers and information managers. These people make their collections easier to use by introducing intermediate information structures in addition to the raw text of the documents. These structures broadly include standardized subject and keyword descriptors and classification systems, together with standardized methods of cataloguing the documents.

Successful publication of collections of documents therefore requires provision of additional intermediate information structures to facilitate the user's task of finding things and getting overviews of the collection. In most cases, it is not economic to employ a professional librarian just as it may not be economic to employ a professional software engineer. The collection is typically managed by someone who understands the domain, with sufficient information technology expertise to operate and configure the software products used. In the same way, the manager of the collection needs to know the basic principles of information science.

This book is aimed at the managers of document collections, and those university students who expect in their professional life to have occasion to manage such collections, without being professional librarians. It is organized in two main parts – the first oriented towards the problem of finding documents in the collection and the second towards the classification systems and controlled vocabularies which not only assist in information retrieval but also make possible an overall view of the collection. It begins by considering the problem as the retrieval of information from a collection of documents, but by the end the problem is seen as the user navigating in an information space.

The information retrieval part opens with an overview outlining the main issues, then proceeds to a discussion of the technologies employed for retrieval and how they are used, then to the information structures used to support information retrieval. Chapter 4 turns to the use of hypertext and the strategy of browsing rather than searching, then to the World Wide Web and the problem of resource discovery in that vast and heterogeneous environment. The information retrieval part concludes with a consideration of documents which are themselves complex structures and the use of the markup language XML to represent and exploit that structure.

Until this point, the text has taken the perspective of the user trying to find information in the document collection. The second part takes the perspective of the managers responsible for the information structures used to organize the collection. It begins with a discussion of classification systems and their basic design principles, then how these classification systems can be used to organize and present the collection. Chapter 9 makes the point that classification systems are designed, not inherent qualities of the documents. This leads to the main design principles, first of classification systems then of subject and keyword systems. The structure part concludes with the use of the structures to support sophisticated visualization of the information spaces.

Two final chapters look at the main issues in creating and managing archives, which are the ultimate source of many published document collections; and general issues of quality, including legal considerations.

The book is intended to assist the reader to learn a specialized technical vocabulary used to describe information-seeking behaviour, the design elements of information spaces and the metrics used in the evaluation of design choices. In particular, the reader will learn to use these concepts in design and criticism of design of information spaces. There are about 75 key concepts, which are gathered into a glossary at the end.

Use of this book as a text is facilitated by exercises and discussion questions at the end of each chapter, and by the inclusion of a sample set of assignments which have been successfully used in teaching this material to a broad group of second-year students at The University of Queensland.

As a university text, this book is relevant to a wide range of disciplines, including journalism, languages, anthropology, management, public administration, law, education, social work, and engineering, as well as more application-oriented information technology students. It is relevant to health informatics, tourism informatics and biological informatics – to any field of activity which generates or requires access to collections of documents.