The Subject Approach to Information

A.C. FOSKETT

FOURTH EDITION

THE SUBJECT APPROACH TO INFORMATION

FOURTH EDITION BY

A C FOSKETT
MA FLA ALAA
Associate Professor
Head of Library Studies
South Australian Institute of Technology

O What a tangled web we weave
When first we practice to retrieve . . .


CLIVE BINGLEY
LONDON

&


LINNET BOOKS
HAMDEN · CONN

Publishing history

FIRST PUBLISHED 1969
REPRINTED 1970
SECOND EDITION PUBLISHED 1971
REPRINTED 1973, 1975, 1976
THIRD EDITION PUBLISHED 1977
REPRINTED 1981

THIS FOURTH EDITION PUBLISHED 1982
BY CLIVE BINGLEY LTD, 16 PEMBRIDGE ROAD, LONDON W11, UK
SIMULTANEOUSLY PUBLISHED IN THE USA
BY LINNET BOOKS, THE SHOE STRING PRESS INC,
995 SHERMAN AVENUE, HAMDEN, CONNECTICUT 06514

SET IN 10 ON 11 POINT PRESS ROMAN BY ALLSET
PRINTED AND BOUND IN THE UK BY
REDWOOD BURN LTD, TROWBRIDGE, WILTS

BINGLEY ISBN: 0-85157-313-4 (HARDBACK)
0-85157-339-8 (PAPERBACK)
LINNET ISBN: 0-208-01934-0 (HARDBACK)
0-208-01970-7 (PAPERBACK)

British Library Cataloguing in Publication Data

Foskett, A. C.
The subject approach to information.—4th ed.
1. Subject cataloguing
I. Title
025.4 Z696
ISBN 0-85157-313-4
ISBN 0-85157-539-6 Pbk

CONTENTS

ILLUSTRATIONS

AUTHOR'S PREFACE

The first edition of this book, published in 1969, was prompted by changes in the Library Association syllabus which for the first time enabled lecturers to treat classification and subject cataloguing as different aspects of the same topic. These changes coincided with the publication of the results of the Cranfield project, which showed clearly that all indexing languages are basically the same. However, although it was possible to teach these new approaches, there was no one textbook which covered the subject approach in the way that I felt was needed. The first edition was intended to fill this gap, and its reception, both in Britain and abroad, showed that it did indeed fulfil a real need. It met with a very positive welcome from lecturers and – more importantly – from students, particularly in the United States, where British textbooks on classification had previously been regarded with something akin to suspicion. (As the Dean of one American library school is alleged to have remarked, '*What* theory of classification?')

The first edition was quickly sold out, and I decided to revise it in the light of comments by colleagues and reviewers, and of developments in my own thinking. My experience in teaching at the University of Maryland during 1968-1969 and summer 1970 was another important factor influencing my decision to revise rather than reprint. I was thus able to include such important developments as the adoption of PRECIS by BNB with the introduction of MARC at the beginning of 1971, and it was a matter of some small pride that my account of the eighteenth edition of Dewey appeared at about the same time as the official publication of the scheme itself, thanks to the cooperation of the editor, Mr Ben Custer, and of BNB, which had started to use DC18 at the beginning of the year, in advance of publication.

The second edition was also well received all over the world, and was accorded the singular honour of translation into Portuguese for use in library schools in Brazil. In due course the question of whether to revise or reprint arose again. My first intention was to prepare the third edition for publication in 1976, to celebrate the centenary of Dewey and Cutter, but a stroke in June 1975 put this out of the question, and

the third edition appeared in 1977. By 1980 I had decided that developments were moving sufficiently rapidly to justify a new edition, revised to take into account the growth in the use of computer-based services, which in 1977 in Australia were a cloud no bigger than a man's hand. This edition largely retains the structure of the third, but the balance has been significantly altered to reflect modern trends.

In the first edition it seemed reasonable to include a chapter on 'the computer' in the section on the future. I debated whether to change this in the second edition, but left it as it was. By 1977, this was obviously the wrong approach, and I brought it forward to chapter 3; however, several reviewers pointed out that very few library school students now do *not* already study computing and computers, so the chapter has been omitted entirely from this edition. It is assumed that students will have sufficient background knowledge in this field not to need further general information. This has enabled me to split the chapter on derived indexing into two parts, one dealing with printed types of derived index, the other with the basics of on-line searching. In edition 3, the whole chapter occupied 16 pages; in edition 4, part 1, chapter 3, occupies 18 pages, part 2, 12, giving nearly twice the coverage. The increase in the size of the chapter on semantics from 15 to 19 pages, and of that on syntax from 29 to 37, are also due to the inclusion of material relating to computer-based systems, while the chapter on the practical aspects, chapter 25, has increased from 16 to 33 pages. Computer-based systems account for almost all of the increase in size of this edition over the third; in fact, only two other chapters have grown, that on the Bibliographic classification, of which the second edition is now well under way, and that on the Dewey decimal classification, where again there is a new edition to describe. Perhaps Dewey gets more than his fair share; but without him, information retrieval in libraries would almost certainly have taken a different course – for better or for worse.

Some reviewers felt that I should have included some kind of philosophical basis for the arguments I present. If there is any kind of philosophy behind my arguments, it is perhaps a behavioural one: what do people actually *do* when trying to find information? From this point of view the chapters in part I do in fact follow a reasonably logical sequence from the simplistic approach of elementary methods of derived indexing to the semantic and syntactic puzzles which face us if we try to delve below the surface. My main excuse must, however, be that of Oliver Edwards, as quoted by Boswell: 'I have tried too in my time to be a philosopher; but, I don't know how, cheerfulness was always breaking in'. This could also serve as a hint to some who seem to take my every word in deadly ernest; although I do at times write with my tongue between my lips (the standard attitude of deep concen-

tration), there are other times when it is equally firmly in my cheek. This text contains its share of what John Martyn so aptly described as 'Alice in Wonderland' citations; irrelevant allusions included by authors 'to add artistic verisimilitude to a bald and otherwise unconvincing narrative', to quote another favourite source. To those who recognize them, I hope they provide a moment's amusement or relaxation; anyone who is momentarily puzzled may safely ignore them.

One of the greatest sources of satisfaction I have gained from the earlier editions is the frequent comment by lecturers and students that the book is *readable*; I hope that I have managed to achieve that primary objective of good technical writing, though I fear that the increasing complexity of the subject matter necessarily makes the presentation harder going. One point is perhaps worth mentioning. The text is intended to be followed through as it stands; as the King of Hearts advised Alice, 'Begin at the beginning, and go on till you come to the end: then stop'. However, some reviewers have indicated that this is not the way that they would use the book as a Text, and I have therefore included a certain amount of repetition to enable those who follow a different route to do so with a minimum of inconvenience. For those who do follow my own path of reasoning, I hope that the repetition will serve as reinforcement of points I believe to be of some importance.

Another point I must stress is that this is a textbook, not a workbook. Familiarity with any of the indexing languages or techniques described must be gained *at first hand*, preferably under the guidance of a tutor, with as much practical work as possible. I have tried to give the background to each topic described, and sufficient detail for basic comprehension, but it is not possible to do more in a work of this kind. Familiarity begets boldness, we are told; in some cases it may also have another well-known effect – 'but the task of filling up the gaps I'd rather leave to you . . . '

I owe a great deal to a multitude of authors, and to colleagues and students at Loughborough Technical College, College of Librarianship Wales, University of Maryland and South Australian Institute of Technology; so many that to name them all would be impossible, while to select a few would be invidious. I hope therefore that they will accept this expression of my sincere thanks as an inadequate but deeply felt acknowledgement of my debt to all of them. I hope I have not committed too many sins of commission; of sins of omission I fear there are many. The author of a work such as this will fully appreciate the aptness of Job's cry: 'My desire is . . . that mine adversary had written a book.' (Job xxxi, 35) There must be few other ways of leaving oneself so vulnerable to the slings and arrows of outrageous (or outraged) critics.

As usual, my wife has been a tower of strength during the gestation

period of this edition; the fact that we are still together after four editions is a source of some amazement to me! Without her constant help and encouragement, I could not have achieved the little that I can claim as mine.

A C FOSKETT
September 1981

LIST OF ABBREVIATIONS

AA 1908	Anglo-American Cataloguing Rules 1908
AACR	Anglo-American Cataloguing Rules 1967, 1978
ABN	Australian Bibliographical Network
ABNO	All But Not Only
ACI	Australian Consolidated Industries
ACICS	ACI Computing Services
AESIS	Australian Earth Sciences Information System
AHCI	Arts and Humanities Citation Index
AID	Associative Interactive Dictionary
AIM-TWX	Abridged Index Medicus – Telex
ALA	American Library Association
ALMS	Automated Library Management System
AMRS	Australian MARC Record Service
ANB	Australian National Bibliography
ANSI	American National Standards Institute
ANTIOPE	French viewdata system
AP	Articulated Prepositional (EPSILON)
ASCA	Automatic Subject Citation Alert
ASCII	American Standard Code for Information Interchange
ASIS	American Society for Information Science
ASSASSIN	Agricultural System for Storage and Subsequent Selection of INformation
ASTIA	Armed Services Technical Information Agency (now DDC)
ATMS	Automated Text Management System
AUSINET	Australian Information Network
BBC	British Broadcasting Corporation
BC	Bibliographic Classification (H E Bliss)
BCA	Bliss Classification Association
BH	Broad Heading (EPSILON)
BLLD	British Library Lending Division
BLR&DD	British Library Research & Development Division
BM	British Museum
BME	Basic Medium Edition (UDC)
BNB	British National Bibliography

BRS	Bibliographic Retrieval Services
BSI	British Standards Institution
BT	Broader Term
BTI	British Technology Index (now CTI)
BUCCS	Bath University Comparative Cataloguing Study
CA	Chemical Abstracts
CAC	Chemical Abstracts Condensates
CAI	Computer Aided Instruction
CAS	Chemical Abstracts Service
CASIA	Chemical Abstracts Subject Index Alert
CATNI	Catchword and Tradename Index (CTI)
CBAC	Chemical-Biological ACtivities
CC	Colon Classification
CDU	Classification Décimale Universelle (= UDC)
CEEFAX	BBC teletext service
CFSTI	Clearinghouse For Scientific and Technical Information (now NTIS)
CIDST	Committee for Information and Documentation in Science and Technology
CIJE	Current Index to Journals in Education
CITE	Current Information Transfer in English
CLA	Canadian Library Association
CLIR	Clearinghouse for Library and Information Resources
CLIS	Clearinghouse for Library and Information Science (now incorporated in CLIR)
CLRU	Cambridge Language Research Unit
CLW	College of Librarianship Wales
CNRS	Centre Nationale pour la Recherche Scientifique
COM	Computer Output Microform/fiche/film
COSATI	Committee On Scientific And Technical Information
CP	Chain Procedure (EPSILON)
CRG	Classification Research Group
CSIRO	Commonwealth Scientific and Industrial Research Organisation
CTI	Current Technology Index (previously BTI)
DC	Decimal Classification (Dewey)
DC&	Decimal Classification Additions, Notes, Decisions
DCD	Decimal Classification Division (Library of Congress)
DDC	Defense Documentation Center
DIANE	Direct Access Information Network for Europe
DISISS	DISsemination of Information in the Social Sciences
DK	Dezimal Klassifikation (= UDC)
DRTC	Documentation Research and Training Centre
ED	ERIC Document

EDRS	ERIC Document Reproduction Service
EE	English Electric Company (now part of GEC)
EJ	ERIC Journal abstract number
EJC	Engineers Joint Council
EMPST	Energy-Matter-Personality-Space-Time
EPSILON	Evaluation of Printed Subject Indexes by Laboratory investigatiON
ERIC	Educational Resources Information Center
EUDISED	EUropean Documentation and Information System in EDucation
FID	Fédération Internationale de Documentation
FID/CCC	Central Classification Committee for UDC
FP	Futility Point
FPC	FP Criterion
GDE	Generalized Data Entry (IBM data base system)
GRACE	GRaphic Arts Composing Equipment
IAC	NASA Industrial Application Center
IBM	International Business Machines Corp
ICI	Imperial Chemical Industries Ltd
ICSU	International Council of Scientific Unions
IEE	Institution of Electrical Engineers
IIB	Institut International de Bibliographie (now FID)
IID	Institut International de Documentation (now FID)
IM	Index Medicus
INSPEC	INformation Service in Physics, Electrotechnology, Computers and control
IR	Information Retrieval
IRE	Institute of Radio Engineers (now Institute of Electrical and Electronic Engineers)
IRRD	International Road Research Documentation
ISBN	International Standard Book Number
ISD	International Subscriber Dialing
ISI	Institute for Scientific Information
ISILT	Information Science Index Languages Test
ITA	Independent Television Authority (UK commercial TV)
KWAC	KeyWord And Context
KWIC	KeyWord In Context
KWOC	KeyWord Out of Context
LA	Library Association (UK)
LAA	Library Association of Australia
LASER	London And South Eastern Region
LASIE	Library Automated Systems Information Exchange
LC	Library of Congress Classification
LCSH	Subject Headings used in . . . the Library of Congress

LEC	London Education Classification
LISA	Library and Information Science Abstracts
LT	Lead Term (EPSILON)
MARC	MAchine Readable Cataloguing
MEDLARS	MEDical Literature Analysis and Retrieval System
MEDLINE	MEDLARS on-line
MeSH	Medical Subject Headings
MIDAS	Multimode International Data Acquisition System
MIT	Massachusetts Institute of Technology
NASA	National Aeronautics and Space Administration
NIH	National Institute of Health
NLL(ST)	National Lending Library for Science and Technology (now part of BLLD)
NLM	National Library of Medicine
NRCd	National Reprographic Centre for Documentation
NSA	Nuclear Science Abstracts
NT	Narrower Term
NTSC	National Television Standards Committee (Color TV system used in USA)
NTIS	National Technical Information Service
OATS	Original Article Tearsheet Service (ISI)
OBAR	Ohio Bar Automated Research system
OBNA	Only But Not All
OCLC	Formerly Ohio College Library Center, now OCLC Inc
OCR	Optical Character Recognition
OECD	Organization for Economic Cooperation and Development
ORACLE	Optical Recognition of Announcements by Coded Line Electronics (ITA teletext service)
P-Note	Provisional extension to UDC
PAL	Colour TV system used in UK and Australia
PC	Prediction Criterion
PCMI	Photo-Chromic Micro-Image
PMEST	Personality-Matter-Energy-Space-Time
PRECIS	PREserved Context Indexing System
PRESTEL	UK Post Office viewdata system
RIE	Resources In Education
RIN	Reference Indicator Number (PRECIS)
RIQS	Remote Information Query System
RS	Rotated String (EPSILON)
RT	Related Term
RT	Rotated Term (EPSILON)
SBA	NASA Small Business Administration Technology Assistance Program

SC	Subject Classification (J D Brown)
SCI	Science Citation Index
SDC	System Development Corporation
SDI	Selective Dissemination of Information
SECAM	Colour TV system used in France and USSR
SHARP	Ships Analysis And Retrieval Project
SIN	Subject Indicator Number (PRECIS)
SLIC	Selective Listing In Combination
SMART	Computer-based retrieval system devised by G Salton
SN	Scope Note
SR	Shunted Relational (EPSILON)
SRC	Standard Reference Code/Standard Roof Classification (BSO)
SSCI	Social Science Citation Index
STAIRS	STorage And Information Retrieval System (IBM)
SYNTOL	SYNTagmatic Organisation Language
Telidon	Canadian viewdata system
TEST	Thesaurus of Engineering and Scientific Terms
TIP	MIT Technical Information Project
UBC	Universal Bibliographic Control
UCLA	University of California at Los Angeles
UDC	Universal Decimal Classification
UKAEA	United Kingdom Atomic Energy Authority
UKCIS	United Kingdom Chemical Information Service
UNISIST	United Nations Information System in Science and Technology
USAEC	United States Atomic Energy Commission (now US Energy Commission)
VDU	Visual Display Unit
VINITI	All-Union Institute for Scientific and Technical Information
VIP	Vocabulary Improvement Project (ERIC)
WLN	Washington Library Network
WRU	Case Western Reserve University

PART 1: THEORY OF
INFORMATION RETRIEVAL SYSTEMS

Chapter 1

INTRODUCTION

Libraries form an essential part of the chain of human communication. Before knowledge was recorded (and even to this day in very primitive societies), individuals formed the repository of knowledge, the bridge between successive generations and between those who generated new information and those who required to use it. The amount of information that can be passed on in this way is limited, and society began to move forward when information of various kinds began to be recorded in relatively permanent forms which could serve as a substitute for the wise man in person. Knowledge only becomes generally useful when it is made available; by recording it, we do our best to ensure that it is permanently available to anyone who may need it, instead of ephemeral and limited to one individual.

Nowadays, the quantity of new information being generated is such that no individual can hope to keep pace with even a small fraction of it, and the problem that we have to face is that of ensuring that individuals who need information can obtain it with the minimum of cost (both in time and in money), and without being overwhelmed by large amounts of irrelevant matter. As W H Auden put it: 'The greatest problem of today is how to teach people to ignore the irrelevant, how to refuse to know things, before they are suffocated. For too many facts are as bad as none at all.'[1]

So, instead of the individual store of knowledge, we have the corporate store: the library; instead of the individual memory, we have the corporate memory: library catalogues and bibliographical tools. And just as the individual whose memory fails him cannot pass on wanted information when it is wanted, so a library whose corporate memory is inadequate will fail in its purpose.

Knowledge and information

We may make a useful practical distinction by defining knowledge and information:

knowledge is what *I* know
information is what *we* know

These pragmatic definitions are in line with the more conventional definitions in the *Concise Oxford dictionary*:

knowledge. Familiarity gained by experience; person's range of information

information. Informing, telling; thing told.

In each case, the fundamental distinction is that knowledge is restricted to the individual who gains it, while information is knowledge shared by communication. If we record knowledge, then it may be communicated at a distance in space and time; we do not have to be face to face with the informant as we do in oral communication. This lends further emphasis to the concept of the library as the repository of recorded knowledge, and the importance of being able to pass on that knowledge to whoever needs it, at the time that it is needed. The problems are compounded by changes in the way that knowledge develops and in the kind of demand that users are likely to place on our information systems, as well as the sheer quantities of knowledge that have accumulated.

The growth of knowledge

In a valuable study carried out for the US Government,[2] Arthur D Little Inc identified three eras of information need. The first of these was the discipline-oriented era, which lasted in effect from the development of printing until well into this century. This was characterized by the division of knowledge into more or less watertight compartments or disciplines; new disciplines grew out of the splitting up, or 'fission', of existing disciplines as particular aspects grew in importance as the result of specialization, and developed into disciplines in their own right. Thus science developed from philosophy; physics developed from science; electricity developed from physics; and electronics developed from electricity. In each case, the new subject represented a fragmentation of the old, but remained within it.

The second era, the problem-oriented era, began to assume importance in the 1930s and particularly in the Second World War. This was characterized by the need to solve particular problems, using whatever disciplines might be necessary, regardless of whether they 'belonged together' or not. A classic example is the invention of the magnetron, a device for generating radio waves at very high frequencies, which formed a vital link in the development of radar. Conventional approaches involved the refinement of existing high frequency devices, but were only successful up to a point, and the solution was only found when a new team of workers, whose background was in the separate, though distantly related, discipline of nuclear physics, became involved and started looking for a solution in the techniques of their own speciality. More recently, we have seen the development of such

subjects as molecular biology, which involves the merging, or 'fusion', of disciplines such as physics and biology which had previously been thought of as totally distinct.

We are now entering the third, or mission-oriented, era, in which demands for information may span a whole range of disciplines. For example, space medicine certainly requires a knowledge of medicine, but in addition involves problems related to space physics, mechanics (the phenomenon of weightlessness), diet, hygiene – the list is formidable. Clearly the old barriers between disciplines, which began to crumble in the problem-orientated era, have now effectively disappeared, which presents further difficulties in the transmission of information. The more remote new knowledge is from an individual's existing range of information, the more difficult it becomes for him to comprehend and incorporate into his own store of knowledge. (Try to envisage explaining the significance of a bid of 'two clubs' in contract bridge to someone who has never seen a pack of playing cards. It is obviously necessary to do a lot of spade work before one can even get to the heart of the question.) The needs of today's information users place demands on our information services far more acute than were normal in past eras, and our information retrieval systems must be adequately developed to meet these demands.

It is perhaps possible to distinguish a fourth era, the polymath era. This preceded the discipline-oriented era, and represents the time when the sum total of human knowledge was sufficiently small for one man to be able to comprehend it all. This era would be of purely historical significance if it were not for the fact that many users of information services seem to assume that the librarian is indeed, as has sometimes been suggested, the last polymath. In today's climate of what might be described as comprehensive specialization, the provision of information services involves teamwork and cooperation on a scale far greater than has been common in the past. No librarian is an island!

In the light of the preceding discussion, we may see that one measure of the success of an information retrieval system is its effectiveness as a means of transmitting information, ie as a means of communication. It is therefore helpful to look at the communication process itself, and the ways in which it is modified by the library environment, but first we must clarify the significance of document surrogates, and pinpoint those aspects of documents which are likely to cause most problems in retrieval.

The importance of catalogues and bibliographies

Libraries contain information in many different physical forms. While for many the book is still the major vehicle for the communication of information, for others the periodical or technical report have taken its

place; for yet others, newer forms such as films or gramophone records are the significant items. It is clear that the same work can appear in several different physical forms: for example, we may have Shakespeare's play *Hamlet* in book form, as a film, or on a record. The intellectual content will be the same in each case, but obviously it is not always practical to try to arrange the different physical forms together. Although it is possible to obtain shelving which does permit this, it is not economical of shelf space, and in some libraries might well be regarded as more of a hindrance than a help.

We cannot therefore rely on the physical arrangement of the items in a library to gather together different versions of the same work; we have to rely on a substitute – a set of records (surrogates) of the content of the library. In addition to the physical form, other factors may influence the place where we choose to keep any given item: we may decide that it should not be removed from the building, so it is placed in the reference section; or it may be suitable for children rather than adults, so it is placed in the children's section; or it may be a rarely used work which is placed in the stack rather than in the section open to the public. All of these factors emphasize the importance of the records, as opposed to the items themselves, for we can gather together in one place the records of items which themselves must perforce be scattered.

Suppose that our library only contains items of one kind, eg books. We can now attempt to arrange these in a way that will be useful, but we still cannot dispense with the records. For any particular book there will be several ways in which we might wish to find it: we may know the author, or the title, or we may need to find it because of its subject. The arrangement of books on our shelves may be by author, in which case we shall be able to find the required book if we know the author, but not if we only know the title or subject; equally, it may be by subject, in which case we shall not be able to find a book if we only know the author. But there is no such restriction when we consider the records we may make, for we can record a book by any and every factor which we think may prove to be of use when we are later searching for it. All we need is a fixed address to which each of these factors will lead us, so that we can locate the physical item no matter how we approach it. Once again, it is the records of the contents of our library which are the essential keys: the corporate memory.

The library's catalogue, however, is only one of the tools which serve as the corporate memory. A library containing large numbers of periodicals will not attempt to list every article in every issue it receives; instead, we rely on indexes, abstracts and similar bibliographical tools, which present the same kind of opportunity – and the same prob-

lems – as the library's catalogue, by enabling us to obtain access to any particular item through a number of different approaches.

Factors which identify

Some of these approaches *identify* the items they refer to. For example, if we state the number of a patent specification, there will only be one item corresponding to that description; if we name an author, we immediately limit severely the number of works which will satisfy us. Title, edition, date of publication, publisher, are all factors of this kind. We can therefore give a definite yes/no answer to an enquiry regarding one of these factors; either we can supply what our reader wants, or we cannot. A reader who asks for a copy of *The tempest* may not mind very much which edition we give him, but he will certainly object if he is given *The white devil*, even though both were published in 1612. An enquiry for AERE Inf/Bib 132 will not be satisfied by AERE Inf/Bib 125. Provided we have entries in our records which will lead us to the place where we can find these items, our search is straightforward.

Factors which do not identify

There are, however, other approaches which do not identify the items which will answer them. If we are asked for a 'nice detective' we have a wide choice of answers, any one of which may satisfy the enquirer. More seriously, if we are asked for something on a particular subject, we may be able to find a number of potential answers from which we shall have to ask our reader to make his own selection. Because authors write from within their own individual nexus of experiences, and readers read within their limits, there will seldom be the exact correspondence that we have with factors which identify; instead, we shall have to try to get as good a match as we can between our reader's needs and what we can supply, accepting that it is unlikely that we shall immediately find the same kind of yes/no answer as is possible when we can identify.

Information retrieval as a form of communication

We may consider information retrieval processes as part of the overall pattern of communication.[3] The most commonly used model of the communication process is that devised by Shannon and Weaver, shown in figure 1a. In this model, we see that a *source* has a *message* which is to be transmitted to a *receiver*; before it can be transmitted, the message must be *encoded* for transmission along the selected *channel*, to be *decoded* before it can be understood by the receiver. In information retrieval, the sources are the originators of the documents we handle; the encoding process includes the choice of the appropriate

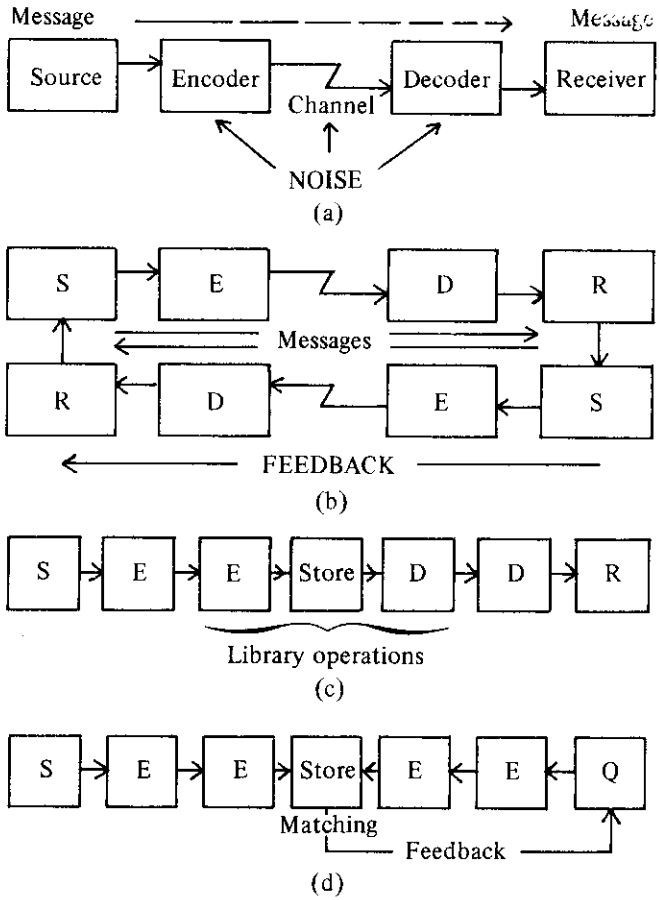

Figure 1: Models of the communication process

(a) The Shannon-Weaver model
(b) Verbal (two-way) communication; involves feedback
(c) The effect of library operations
(d) The query situation: feedback through the matching process

words and their translation into print (or whatever medium is used); the channel is the document and its progress from originator to user; and the decoding process involves the user and his ability to comprehend the message in the form in which it is presented to him. The final element in the model is *noise*. Noise may be defined as anything which detracts from the fidelity of the transmission of the message from source to receiver. Shannon and Weaver were concerned with the transmission of messages over telephone wires, but the concept of noise, which is obviously relevant to that situation, can be generalized to cover all kinds of interference with communication – for example, unwanted documents retrieved in answer to a request.

Two other useful concepts are *entropy* and *redundancy*. Entropy is a measure of the degradation or disorganization of the universe, and may be used to refer to the degradation of a message by noise; there is a mathematical sense in which it is a measure of the rate of transfer in a message. Redundancy is a measure of the amount of superfluous material in a message which can be omitted without loss of significance, and as such would appear to be undesirable. However, when we take into account the losses experienced through noise in its various manifestations, we can see that a message containing a certain amount of redundancy may well still be comprehensible, whereas one which starts off with no redundancy may well be indecipherable as a result of relatively small amounts of noise.

If we consider normal verbal communication, we can see that the model (figure 1b) is in fact that of figure 1a doubled, so that the original source functions also as a receiver and vice versa. In this situation, a further important element enters the picture: the idea of *feedback*. If the message becomes distorted on its way from source to receiver, the receiver can immediately query it (I didn't quite catch that? or, Could you explain that again, please? or *What?*) Feedback can thus reduce the effect of noise and improve the fidelity of communication.

Unfortunately, when we are dealing with documents, there can be little, if any, feedback from user to originator, and any that does take place is usually at a distance, for example through correspondence. The source has no control over who receives his messages, and cannot therefore direct them to a specific audience, though he may well have a fair idea of the *likely* audience. The receivers in their turn cannot be certain that they have understood the messages correctly, or that they have located the messages they are looking for. Librarians are all too familiar with the user who returns a book with the comment that it was not quite what was wanted. Indeed, the situation is in some ways made more difficult by the interposition of additional encoding processes in the library, which then require further decoding processes

on the part of the user. We normally arrange books by putting a code on the spine: a class mark; and we identify them in our catalogues by a further series of codes: catalogue entries (figure 1c). Many readers have difficulty in comprehending these codes; some indeed regard them as a barrier imposed by librarians to prevent the users from obtaining easy access to the items they want and thus ensuring a measure of job security. Catalogue entries and class marks can in fact give a great deal of information, but there is no doubt that they do form an additional complication in the chain of communication, and as such are an additional source of noise.

Furthermore, as may be seen from the diagram, we have introduced yet another complication. Not only is transmission of the message indirect; it is also *delayed* by being placed in a *store* of some kind. A book may be regarded as such a store, as may any kind of document, but we are primarily concerned here with the library, as represented in catalogues and bibliographies, as our store. In this context, such relatively trivial matters as misshelving of documents or misfiling of catalogue cards all add their quota of noise to the communication process.

Searching

So far we have considered the passive approach: the library provides material, which the readers select in a fairly indeterminate way. However, much of the use made of libraries is active: readers come to the library seeking information on particular subjects, and expect our systems to be able to provide the answers. In this situation, the receiver becomes a source, encoding a message in the form of an enquiry. We now have to discover any messages in our store which appear to match the enquiry; having found some, we can pass them on to the enquirer, who can decide whether they match his needs. In the light of our response, the enquirer may modify his message in an attempt to achieve a closer match with his requirements; in other words, we have a degree of feedback in the system, which may enable us eventually to satisfy a request despite an initial failure. This failure may arise from a variety of causes: the enquirer may not be able to express himself clearly, or he may not be very sure exactly what it is he needs (if he knew the answer he would not need to ask the question!); or our encoding processes may be inadequate; or the original source (the author) may not have made his message clear, or may even have had some quite different message in mind (for example, the answer to a question from a librarian on the optimum number of staff required to man the circulation desk may be found in a book on supermarket management). The significance of this feedback mechanism is discussed further in chapter 2 in relation to iterative and heuristic searching.

It is helpful to bear this model in mind when considering any information retrieval system. Are there any factors which tend to increase noise? Which is the most appropriate channel for any particular message? Libraries now take the whole of recorded knowledge as their province, not merely the more conventional books and periodicals; perhaps a tape-slide presentation or a videotape may be the best answer to an enquiry. How accurate are our encoding and decoding processes? How much feedback can be built into the system? With these points in mind we can be more confident of our ability to achieve the match between our readers' needs and what we can supply which we have already referred to. The study of communication has shown that it is quite rare for 100% success to be obtained even under optimum conditions, while the plays of such authors as Harold Pinter show clearly the extent to which we all too often fail to achieve any kind of communication at all. With this in mind, we should strive to optimize our information retrieval systems, while recognizing that they are always likely to remain imperfect.

Some further factors to be considered

There is a further set of factors which do not identify particular documents but which are very important in selecting once we have identified potentially useful items. These include physical form – we may have the very thing, but only as a microfiche, useless without a reader; language – just what is wanted, but in Russian; level – we want an elementary introduction, but only have a doctoral thesis; intended audience – a book intended for one particular group of people (eg fluid mechanics for civil engineers) may not be well suited to the needs of another group; author's viewpoint – a Marxist history of Christianity may offend the practising churchgoer. It will be clear that we can only consider these factors when we have selected a set of documents which appear to match the reader's needs as far as their subject content is concerned.

This book is concerned with a discussion of the problems of optimizing our responses to requests for information on subjects. This is not to suggest that identifying factors such as authors' names do not present any problems; the fact that it took some twenty years of discussions to produce a new edition of the Anglo-American code,[4] which has since been the subject of continuous further discussion, shows very plainly that they do! The problems of subject approach to information, however, are more severe because they are more indeterminate; we never reach the stage of being able to say we have finished a search conclusively. A great deal of research has been done on these problems; much more remains to be done. This book is an attempt to show the present state of the art, in a way which will be acceptable as an ele-

mentary textbook; it does not pretend to be an advanced study, of which there are many, but rather to give beginners some understanding of present theories and ideas.[5]

Bibliography

1 Auden, W H, quoted by Osborne, C: *W H Auden: the life of a poet.* Harcourt, Brace, Jovanovich, 1979.

2 Arthur D Little, Inc: *Into the information age: a perspective for federal action on information*, prepared by Arthur D Little Inc for the National Science Foundation. American Library Association, 1978.

The information society: issues and answers, edited by E J Losey. Oryx Press, 1978.

3 Berlo, D K: *The process of communication: an introduction to theory and practice.* Holt, Rinehart and Winston, 1960. There are a number of books on communication, but Berlo's approach is peculiarly well suited to many of the problems of information retrieval. Chapters 1, 2, 3, 7 and 8 are particularly useful.

4 *Anglo-American cataloguing rules*; prepared by the American Library Association, the Library of Congress, the Library Association and Canadian Library Association. Revised edition, 1978. (AACR2) This edition resolved the differences still outstanding in the 1967 edition between British and American practice; these differences were first stated in the *Anglo-American cataloguing code* of 1908, some 70 years earlier.

5 Of the making of many books on information retrieval there is no end. The following selection is listed alphabetically by author; 'elementary' indicates a work that can be read with profit by the novice; 'intermediate' indicates a work which should be read, but preferably after some preliminary grounding has been obtained; 'advanced', works which may be read by the student wishing to pursue the subject beyond the level of this text.

Artandi, S *ed*: 'Rutgers series on systems for the intellectual organization of information'. This series includes volumes on UDC, SYNTOL, CC, faceted classification, coordinate indexing, and the alphabetical approach; all of them are of high standard and contain much original material. Advanced.

Austin, D: *PRECIS: a manual of concept analysis and subject indexing.* Council of the British National Bibliography, 1974. Intermediate.

Bakewell, K G B: *Classification and indexing practice.* Bingley, 1978. Intermediate.

Brown, A G: *An introduction to subject indexing.* 2v. Bingley, 1976. Elementary. (Programmed text)

Buchanan, B: *Theory of library classification*. Bingley, 1979. Elementary.

Coates, E J: 'Classification in information retrieval: the twenty years following Dorking.' *Journal of documentation, 34* (4) December 1978, 288-299.

Coates, E J: *Subject catalogues: headings and structure.* Library Association, 1960. One of the few worthwhile books on the alphabetical approach. Unfortunately, this excellent work is now out of print, and it does not of course include recent developments such as PRECIS.

Foskett, D J: 'Classification'. (Chapter in Batten, W E *ed*: *Handbook of special librarianship*. Aslib, fourth edition 1975). A concise but useful summary. Intermediate.

Foskett, D J: *Classification and indexing in the social sciences.* Butterworths, second edition, 1975. Discusses many general problems as well as those specific to the social sciences. Intermediate.

Foskett, D J: *Classification for a general index language*. Library Association, 1970. (LA Research pamphlet no 2.) An exposition of the theoretical development of the new general classification scheme being prepared by the CRG, this is of wider interest also as a statement of the problems to be dealt with. Elementary.

Heaps, H S: *Information retrieval: computational and theoretical aspects.* Academic Press, 1978. Advanced.

Herdman, M M: *Classification: an introductory manual*; revised by J Osborn. Third edition. American Library Association, 1978. Elementary.

Hutchins, W J: *Languages of indexing and classification*. Peter Peregrinus, 1975. A linguistic study. Advanced.

Key papers in the design and evaluation of information systems, edited by D W King. Knowledge Industry Publications for the American Society for Information Science, 1978.

Kochen, M: *Principles of information retrieval*. Melville, 1974. Advanced.

Lancaster, F W: *Information retrieval systems: characteristics, testing and evaluation*. Second edition. Wiley, 1979. Intermediate.

Lancaster, F W: *Vocabulary control for information retrieval*. Information Resources Press, 1972. Intermediate.

Langridge, D: *Approach to classification*. Bingley, 1973. Elementary.

Langridge, D: *Classification and indexing in the humanities*. Butterworths, 1976. Intermediate.

Maltby, A: *Classification in the 1970's: a second look.* Bingley, 1976. A revision of a useful work, giving a variety of opinions on current developments. Intermediate.

Needham, C D: *Organizing knowledge in libraries: an introduction*

to cataloguing and classification. Deutsch, second edition 1971. Useful though most of the emphasis is on cataloguing. Elementary.

Palmer, B I: *Itself an education.* Library Association, second edition 1971. The six lectures on various aspects of classification which formed the first edition published in 1962 are reprinted, together with a valuable new chapter on PRECIS by Derek Austin. Elementary.

Sayers, W C Berwick: *Manual of library classification.* Fifth edition revised by A Maltby, Deutsch, 1975. The best presentation of the traditional approach to classification, this edition has been thoroughly revised to introduce modern theory. Elementary.

Sharp, J: 'Information retrieval'. (Chapter in Batten, W E, *ed: Handbook of special librarianship.* Aslib, fourth edition 1975.) Like the chapter on classification, this is a concise but valuable summary. Intermediate.

Van Rijsbergen, C J: *Information retrieval.* Second edition. Butterworths, 1979. Advanced.

Vickery, B C: *Classification and indexing in science and technology.* Butterworths, third edition 1975. Of more general scope than its title implies, this work, like Coates, is not easy, but is essential reading.

Vickery, B C: *Information systems.* Butterworths, 1973. Advanced. (In effect, this might be regarded as the third edition of *On retrieval system theory*, in that it attempts to set out a unified theory of information retrieval, taking account of developments since the second edition in 1965.)

Vickery, B C: *On retrieval system theory.* Butterworths, second edition 1965. One of the most satisfactory attempts to present a unified theory covering all IR methods. Advanced.

Vickery, B C: *Techniques of information retrieval.* Butterworths, 1970. Emphasizes the relationship between information retrieval and other aspects of library science, particularly bibliography. Intended as a textbook, it contains many valuable examples. Intermediate.

Wellisch, H: *Subject retrieval in the seventies: new directions.* Proceedings of an international symposium . . . 1971, edited by Hans Wellisch and T D Wilson. Greenwood Publishing Company, 1972 (University of Maryland School of Library and Information Services: Contributions in librarianship and information science number 3). Intermediate.

For developments over the past few years it is helpful to consult the chapters on classification in *British librarianship and information science, 1966-1970* and *1971-1975.* These also have comprehensive bibliographies. Useful series of articles include 'The year's work in cataloging and classification' in *Library resources and technical services*, and the 'state-of-the-art' reviews published in the *Journal of documentation* under the general heading 'Progress in documentation' eg

'Theories and models in information retrieval' and 'Information concepts for information science'. The Aslib series *Informatics* consists of the proceedings of conferences held by Aslib groups concerned with various aspects of information processing and retrieval. The relevant chapters in the *Annual review of information science and technology* are also valuable for the student wishing to pursue the subject in depth.