

INDEXING AND ABSTRACTING
IN
THEORY AND PRACTICE



INDEXING AND ABSTRACTING IN THEORY AND PRACTICE

F. W. LANCASTER



The Library Association
London

© F.W. Lancaster, 1991

Published by
Library Association Publishing Ltd.
7 Ridgmount Street
London WC1E 7AE
and by
The Graduate School of Library and
Information Science
University of Illinois

All rights reserved. No part of this publication may be photocopied, recorded or otherwise reproduced, stored in a retrieval system or transmitted in any form or by any electronic or mechanical means without the prior permission of the copyright owner and publisher.

First published 1991

British Library Cataloguing in Publication Data

Lancaster, F.W. (Frederick Wilfrid)

Indexing and abstracting in theory and practice

1. Indexing

1. Title

025.48

ISBN 1-85604-004-6

Designed by Bob Chapdu, IntelliText Corp., Champaign, IL 61826, USA. Typeset in Garamond Book by the Publications Office, Graduate School of Library and Information Science, University of Illinois. Printing and binding by Cushing-Malloy Inc., Ann Arbor, MI 48107, USA.

PREFACE

MY INTEREST IN INDEXING and abstracting dates back to 1957-1959 when I served as editor of an in-house abstracts bulletin produced by Tube Investments Limited, Birmingham, England. The main stimulus for producing the present book occurred in 1986. I was asked to prepare a course in indexing and abstracting for the Arab League Documentation Centre and discovered that no book then available dealt with the subject in the way I wanted to present it. Good books on subject indexing did exist but most were either out-of-date or espoused one particular indexing method. Excellent books on abstracting were (and are) available and this work should be considered to complement these rather than to replace them. While a few books endeavored to treat indexing and abstracting together, I felt that they did not cover the subject comprehensively and, in any case, tended to deal with the two topics as separate activities, whereas I wanted to stress their similarities rather than their differences.

This book, then, is intended primarily as a text to be used in teaching indexing and abstracting in schools of library and information science. Nevertheless, I hope it may be of some value to all individuals and institutions involved in information retrieval and related activities, including librarians, managers of information centers, and database producers.

Indexing is closely related to the subject of vocabulary control, but this topic is not dealt with in detail here because it is covered in my book *Vocabulary Control for Information Retrieval*. While I concentrate on indexing and abstracting as practiced by published indexing and abstracting services (in paper or electronic form), rather than library catalogs, the principles are the same, and I hope that the book may be of interest and value to those concerned with improving subject access in online catalogs.

At various stages in the preparation of this text I received considerable help from three graduate assistants, Jill Byttner, Lorraine Haricombe and Beverly Rauchfuss, and I must offer my sincere thanks to them, and also to Kathy Painter, who put the text into machine-readable form.

F. W. Lancaster
University of Illinois
Urbana, April 1990

CONTENTS

Preface	ix
List of exhibits	x
Part 1. Theory and Description	
Chapter 1. Introduction	1
Chapter 2. Indexing Principles	5
Chapter 3. Indexing Practice	19
Chapter 4. Pre-Coordinate Indexes	41
Chapter 5. Consistency of Indexing	60
Chapter 6. Quality of Indexing	74
Chapter 7. Abstracts: Types and Functions	86
Chapter 8. Writing the Abstract	97
Chapter 9. Evaluation Aspects	116
Chapter 10. Approaches Used in Indexing and Abstracting Services	138
Chapter 11. Enhancing the Indexing	169
Chapter 12. On the Indexing and Abstracting of Imaginative Works	182
Chapter 13. Natural Language in Information Retrieval	193
Chapter 14. Automatic Indexing, Automatic Abstracting and Related Procedures	219
Chapter 15. The Future of Indexing and Abstracting Services	247
Part 2. Practice	
Chapter 16. Indexing Exercises	261
Chapter 17. Abstracting Exercises	276

Appendices

Appendix 1 286

Appendix 2 288

Appendix 3 291

References 293

Index 314

LIST OF EXHIBITS

1 The role of indexing and abstracting in the larger information retrieval picture	2
2 The problem of retrieving pertinent items from a database	3
3 Effect of record length on retrievability	7
4 Example of item indexed from various points of view	9
5 Conceptual analysis translated into three controlled vocabularies	18
6 The two indexing dimensions of a document	25
7 Diminishing returns in indexing	25
8 Information retrieval system represented as a matrix	31
9 Index form as used by the National Library of Medicine in 1989	33
10 Typical Mooers' indexing form	35
11 Portion of a specialized vocabulary on digital computers	36
12 Section of microthesaurus of the Air Pollution Technical Information Center	37
13 Specimen entries from a published entry vocabulary	39
14 Entries for a SLIC index	43
15 Entries for an index based on systematic cycling (<i>Excerpta Medica</i> model)	45
16 Sample entries from a KWIC index	46
17 Sample entries from a KWOC index	49
18 Alternative format for a KWOC index	50
19 Specimen entries from the <i>British Technology Index</i>	55
20 Farradane's system of relations	57
21 Terms (A-J) assigned to the same document by five different indexers	61

22	Possible factors affecting consistency of indexing	62
23	Relationship between consistency and number of terms assigned	63
24	Effect of number of terms assigned on indexer consistency (two indexers)	64
25	Two approaches to indexing an article entitled <i>When Bystanders Just Stand By</i>	69
26	Two approaches to indexing an article entitled <i>A Children's Literature Course for Parents</i>	70
27	Two approaches to indexing an article entitled <i>Mentoring in Graduate Schools of Education</i>	71
28	Two approaches to indexing an article entitled <i>Closed Captioned Television: A New Tool for Reading Instruction</i>	72
29	Differences in conceptual analysis for an article entitled <i>The Disappearing Act: a Study of Harlequin Romances</i>	72
30	Factors affecting the results of a search in a database	75
31	Example of important item missed by simple indexer omission	77
32	Factors that may affect the quality of indexing	79
33	Indexer consistency related to user interests	82
34	Indicative abstract	88
35	Informative abstract	89
36	Example of a critical abstract	90
37	Structured abstract	91
38	Modular abstracts	92
39	Modular index entries	93
40	Comparison of mini-abstract, author's summary, and abstracts from <i>Chemical Abstracts</i> and <i>Biological Abstracts</i>	94-95
41	Hints for abstractors	99
42	Abstracting principles published by the Defense Documentation Center (1968)	102

43	Essentials of abstracting	105
44	Hypothetical results from a test of relevance predictability	106
45	Rules for abstractors that relate to retrievability characteristics of abstracts	112
46	Growth of the science literature on AIDS	121
47	AIDS literature: coverage by language, 1982-1987	122
48	AIDS literature: coverage by country, 1982-1987	122
49	Number of journals that published articles on AIDS, 1982-1987	122
50	Scatter of the journal literature on AIDS	124
51	Plot of the scatter of the AIDS literature	125
52	Science journals that published the most papers on AIDS, 1982-1987	125
53	Hypothetical example of distribution of "superconductor" items under terms in a printed index	128
54	Distribution of items on cellular immunology in the pig under terms in <i>Index Medicus</i>	128
55	Scatter of items under index terms	131
56	Example of entries from <i>Cumulated Index Medicus</i>	140
57	Examples of entries from <i>Medical Subject Headings</i>	141
58	Examples from <i>Medical Subject Headings</i> tree structures	142
59	Sample entries from author index to <i>Cumulated Index Medicus</i>	143
60	Sample entries from the <i>Applied Science and Technology Index</i>	144
61	Sample entries from the <i>Engineering Index Monthly</i> (October 1986)	145
62	Sample entries from author index to the <i>Engineering Index</i>	146
63	Sample entries from <i>Library and Information Science Abstracts</i>	147

64	Sample entries from the subject index to <i>Library and Information Science Abstracts</i>	148
65	Subject categories used by <i>Biological Abstracts</i>	150
66	Sample entries from <i>Biological Abstracts</i>	151
67	Sample entries from subject index to <i>Biological Abstracts</i>	152
68	Sample entries from subject index to <i>Chemical Abstracts</i>	153
69	Sample entries from keyword index to <i>Chemical Abstracts</i>	154
70	Sample entries from the formula index to <i>Chemical Abstracts</i>	155
71	Sample abstracts from <i>Sociology of Education Abstracts</i>	156
72	Sample of subject index entries from <i>Sociology of Education Abstracts</i>	157
73	Sample entries from an index of the type used in the Excerpta Medica series	158
74	Sample entries from the <i>Current Technology Index</i>	159
75	Sample PRECIS entries from the <i>British Education Index</i>	61
76	Sample entries from the <i>Social Sciences Citation Index</i>	162
77	Sample entry from the Source Index to the <i>Social Sciences Citation Index</i>	162
78	Sample entry from the Permuterm Subject Index to the <i>Social Sciences Citation Index</i>	163
79	Sample page from <i>Current Contents</i>	164
80	Sample entries from the keyword index to <i>Current Contents</i>	165
81	The EJC system of role indicators	174
82	Semantic infixes in the Western Reserve system	176
83	Role indicators of the Western Reserve system	177
84	Telegraphic abstract as recorded on tape	178
85	Precision devices create smaller classes; recall devices	

build larger ones	181
86 Structure of the Pejtersen scheme for the indexing of fiction	188
87 Example of a novel indexed using the Pejtersen approach	188
88 Two possible synopses for <i>The Story of Peter Rabbit</i> by Beatrix Potter	189
89 Example of an entry from <i>Masterplots</i>	190-191
90 Comparison of abstract and indexing using a controlled vocabulary	200
91 The pros and cons of free text versus controlled vocabulary	202
92 Example of entry in the TERM database	217
93 The essential problems of information retrieval	220
94 Example of thesaurus entries derived by automatic methods	233
95 Citation/reference linkages	234
96 Example of a Luhn auto-abstract	237
97 Example of an extract produced by the ADAM automatic abstracting system	239
98 Filtering levels in a paperless publishing environment	260

INTRODUCTION

THE MAIN PURPOSE of indexing and abstracting is to construct *representations* of published items in a form suitable for inclusion in some type of *database*. This database of representations could be in printed form (as in an indexing/abstracting publication such as *Chemical Abstracts* or the *Engineering Index*), in machine-readable form (in which case the database will often be roughly equivalent to a printed service), or in card form (as in a conventional library catalog).

The role of the indexing/abstracting operations within the larger framework of information retrieval activities in general is illustrated in Exhibit 1. First, the producer of the database selects from the population of newly published documents those that meet certain criteria for inclusion in the database. The most obvious criterion is the subject dealt with, but others, such as type of document, language, or source, may also be important. For those databases that deal primarily with articles from journals, the selection criteria will usually focus on the journal rather than the article; that is, certain journals will be covered and others not (although some journals may be indexed in their entirety and others selectively). To a large extent the coverage of many databases is governed by considerations of cost-effectiveness. Particularly in the case of databases dealing with a highly specialized field, only those journals that publish most on the subjects of interest will be included.

The items selected for inclusion in the database must be "described" in various ways. Descriptive cataloging procedures (not explicit in Exhibit 1) identify authors, titles, sources, and other bibliographic elements, indexing procedures identify the subject matter dealt with, and abstracting may be used to summarize the contents of the item. The terms used in indexing will frequently be drawn from some form of controlled vocabulary, such as a thesaurus (the "system vocabulary" of Exhibit 1), but may instead be "free" terms (e.g., drawn from the document itself). These description activities create document representations in a form suitable for inclusion in the database. The documents themselves will usually go to a different type of database (document store) such as the shelves of a library.

Members of the community to be served use the database primarily to satisfy various information needs. To do this they must convert an information need into some form of "search strategy," which may be as simple as selecting a single term to consult in a printed index or card catalog, or may involve the combining of many terms into a more elaborate

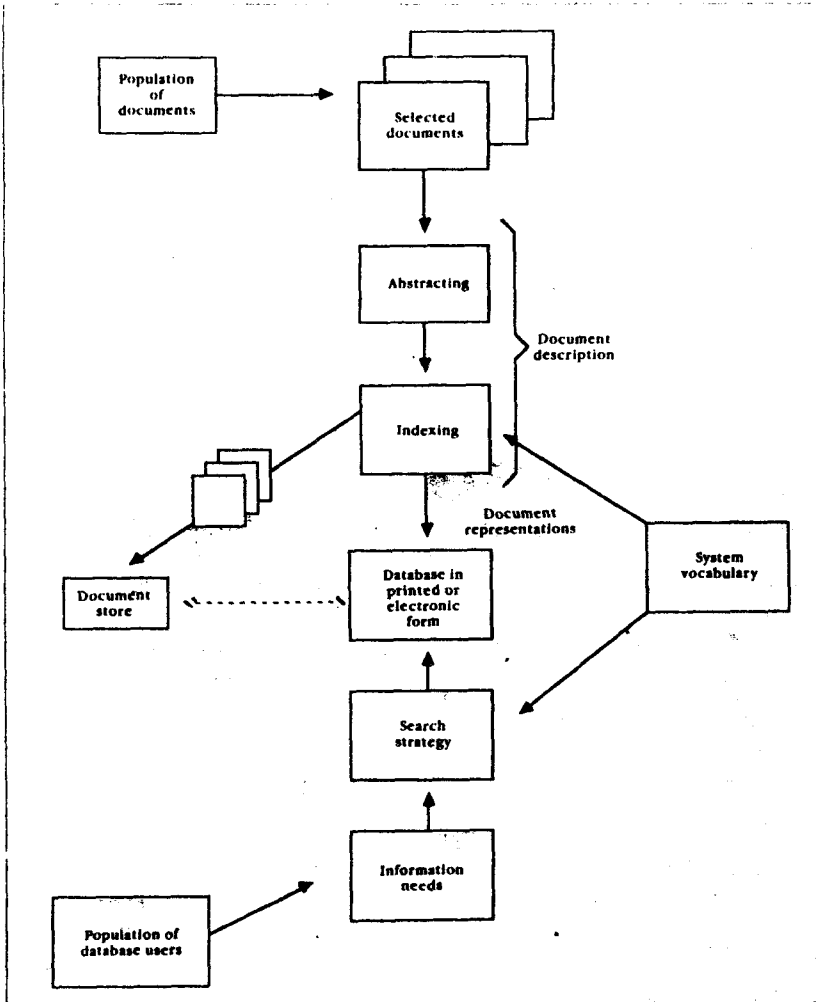


Exhibit 1

The role of indexing and abstracting in the larger information retrieval picture

and sophisticated strategy used to interrogate a database through a computer terminal.

In searching a database, of course, one wants to find items that are useful in satisfying some information need, and to avoid retrieving items that are not useful. Terms such as "relevant" and "pertinent" are frequently used to refer to "useful" items, and these terms have been defined in several different ways. There is a lot of disagreement as to what "relevance" and "pertinence" really mean (Lancaster, 1977). In this book I will consider as synonymous the expressions "useful," "pertinent," and "relevant to an information need." That is, a pertinent (useful) item is one that contributes to the satisfaction of some information need.

The information retrieval problem is depicted graphically in Exhibit 2. The entire rectangle represents a database and the items it contains. The plus (+) items are those that a hypothetical requester would find useful in satisfying some current information need, and the minus (—) items are those that he would judge not useful. For any particular information need there will be many more — items than + ones. Indeed, if the diagram were drawn "to scale," one would expect that the 11 useful items might be accompanied by a whole wall of useless ones. The problem is to retrieve as many as possible of the useful items and as few as possible of the useless ones.

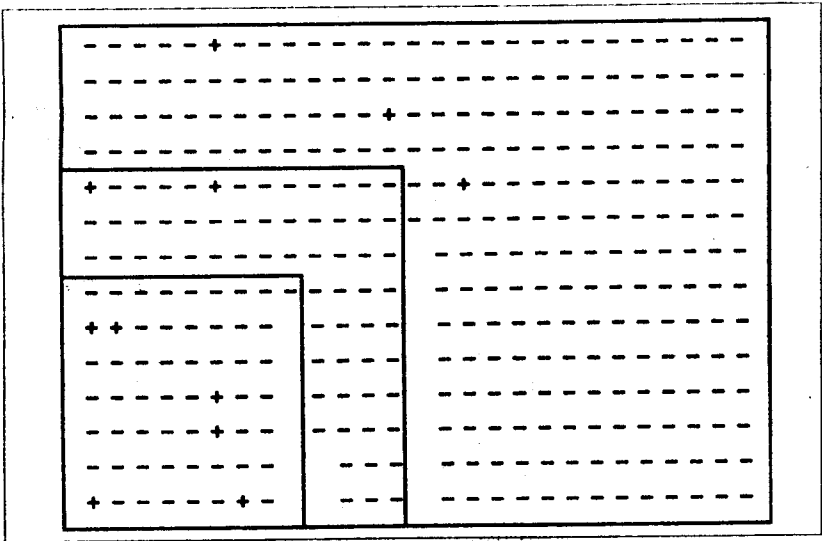


Exhibit 2
The problem of retrieving pertinent items from a database

The smaller of the two interior rectangles of Exhibit 2 represents the results of a search performed in the database. It retrieved 57 items, of which 6 were useful and 51 not useful. The ratio of useful items to total items retrieved ($6/57$ or about 10 percent in this case) is usually referred to as a *precision ratio*. The ratio commonly used to express the extent to which all the useful items are found is the *recall ratio*. In this case the recall ratio is $6/11$ or about 54 percent.

To improve recall in this situation one would probably need to search more broadly. This is depicted in the larger of the two interior rectangles. Through searching more broadly, recall has been raised to $8/11$ (73 percent)), but precision has declined further to $8/112$ or about 7 percent. It is an unfortunate characteristic of the information retrieval situation that an improvement in recall will usually cause a deterioration in precision and vice versa.

Exhibit 2 suggests another phenomenon. It might be possible to search sufficiently broadly to find all the useful items (i.e., achieve 100 percent recall), but precision would probably be intolerable. Furthermore, the larger the database the less tolerable will be a low precision. While a user might be willing to look at abstracts of, say, 57 items to find 6 useful ones, he may be much less willing to examine 570 abstracts for 60 useful ones. With very large databases, then, it becomes increasingly difficult to achieve an acceptable level of recall at a tolerable level of precision.

In this book I use the term *recall* to refer to the ability to retrieve useful items and *precision* to refer to the ability to avoid useless ones. There are other measures of the performance of searches in a database (see, for example, Robertson [1969]), some more mathematically exact, but *recall* and *precision* give the general picture and still seem to be the obvious measures to use to express the results of any search that simply divides a database into two parts (retrieved and not retrieved).*

It is clear from Exhibit 1 that many factors determine whether a search in a database is successful or not. These include the coverage of the database, indexing policy, indexing practice, abstracting policy and practice, the quality of the vocabulary used to index, the quality of the search strategies, and so on. This book makes no attempt to deal with all of these factors (although they are all interrelated) but focuses on the important activities of document description or, at least, those concerned with the content of documents.

*A search that ranks output in order of "probable relevance" requires a somewhat different measure that in effect compares the ranking achieved with some ideal ranking.

INDEXING PRINCIPLES

WHILE THE TITLE of this book refers to "indexing," the scope is actually restricted to subject indexing and to abstracting. Subject indexing and abstracting are closely related activities in that both involve preparing a representation of the subject matter of documents. The abstractor writes a narrative description or summary of the document, while the indexer describes its contents by using one or several index terms, usually selected from some form of controlled vocabulary.

The main purpose of the abstract is to indicate what the document is about or to summarize its contents. A group of index terms can serve the same purpose. For example, the following set of terms gives a fairly good idea of what is dealt with in some hypothetical report:

- Information Centers
- Resource Sharing
- Union Catalogs
- Cooperative Cataloging
- Online Networks
- Interlibrary Loans

In a sense, such a list of terms can be considered to act as a kind of mini-abstract. It would serve such a purpose if all the terms were listed together in a published index or were printed out to represent an item retrieved from some database as a result of a computer search.

More obviously, the terms assigned by an indexer serve as access points through which a bibliographic item can be located and retrieved in a subject search in a published index or machine-readable database. Thus, in a printed index, one should be able to find the hypothetical item mentioned earlier under any one of the six terms. In a computer-based retrieval system, of course, one would expect to be able to find it under any one of these terms or, indeed, any combination of them.

The distinction between indexing and abstracting is becoming increasingly blurred. On the one hand, a list of index terms can be printed out to form a mini-abstract. On the other, the text of abstracts can be stored in a computer-based system in such a way that searches can be

performed on combinations of words occurring in the text. Such abstracts can be used instead of index terms, to allow access to the items, or can supplement the access points provided by the index terms. To some extent this changes the role of the abstractor, who must now be concerned not only with writing a good clear description of the contents of a document but also with creating a record that will be an effective representation for retrieval purposes.

If indexing and abstracting were looked upon as fully complementary activities, the character of the indexing operation might change somewhat. For example, the indexer could concentrate on assigning terms that supplement the access points provided in the abstract. However, such complementarity must be fully recognized and understood by the user of the database. Otherwise a set of index terms alone would give a very misleading picture of the content of an item.

Length of Record

One of the most important properties of a representation of subject matter is its length. The effect of record length is illustrated in the example of Exhibit 3. At the left are various representations of the content of a journal article in the form of narrative text; at the right are two representations in the form of lists of index terms.

The title is a general indication of what the article is about. The brief abstract gives more detail, indicating that survey results are presented in the article and identifying the major questions addressed. The extended abstract carries this further, identifying all of the survey questions and giving the size of the sample used in the study.

The more information given, the more clearly the representation indicates the scope of the article and the more likely it is to indicate to a reader whether or not it satisfies some information need. For example, one might be looking for articles mentioning U.S. attitudes toward various Arab leaders. The title gives no indication that this specific topic is discussed and the brief abstract, by focusing on other topics, suggests that it may not be. It is only the extended abstract that shows that the article includes information on this subject.

The longer the representation, too, the more access points it provides. If title words were the only access points, this item would probably be missed in many searches for which it might be considered a valuable response. As the length of the representation is increased, so is the