Information Retrieval Design

Principles and Options for Information Description, Organization, Display, and Access in Information Retrieval Databases, Digital Libraries, Catalogs, and Indexes

by

James D. Anderson
and
José Pérez-Carballo

Ometeca Institute
ciencia y humanidades • humanities & science • ciência e humanidades
PO Box 12109, St. Petersburg, FL 33733-2109 USA
Distributed by University Publishing Solutions
302 Ryders Lane, East Brunswick, NJ 08816
732-220-1211, fax 732-418-1921, www.upublishing.com

2005

Ometeca Institute

Published in the United States of America
by Ometeca Institute, Inc.
P.O. Box 12109, St. Petersburg, FL 33733-2109

The paper used in the cloth library edition meets the minimum requirements of American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI/NISO Z39.48 – 1992.
Manufactured in the United States of America.

Cover art by David A. Haywood, the Pennsylvania State University.
Cover design by Jon Hansen, University Publishing Solutions, LLC.

Cataloging in publication, prepared by James D. Anderson

Anderson, James D. (James Doig), 1940-

Information retrieval design : principles and options for information description, organization, display, and access in information retrieval databases, digital libraries, catalogs, and indexes / by James D. Anderson and José Pérez-Carballo. – St. Petersburg, Fla. : Ometeca Institute ; East Brunswick, N.J. (302 Ryders Lane, East Brunswick, NJ 08816) : Distributed by University Publishing Solutions, LLC, 2005.

xiv, 617 p. : ill. ; 28 cm.

Includes glossary (p. 543-565), bibliography (p. 567-590) and index (p. 591-616).

ISBN 1-9763547-0-5 (trade paper)

ISBN 1-9763547-1-3 (cloth library edition)

ISBN 1-9763547-2-1 (ebook)

I. Pérez-Carballo, José. II. Title. 1. Information retrieval. 2. Indexing. 3. Database design.

Tables of Contents

Brief Table of Contents

Prefatory material : 1

Part I. Chapter 1. Introduction and Background Issues : 7

Part II: Design Decisions : 43

Chapter 2. Subject Scope and Domain : 47

Chapter 3. Documentary Scope : 73

Chapter 4. Documentary Domain : 87

Chapter 5. Display Media : 91

Chapter 6. Documentary Units : 103

Chapter 7. Indexable Matter : 111

Chapter 8. Analysis and Indexing Methods : 117

Chapter 9. Exhaustivity : 177

Chapter 10. Specificity : 185

Chapter 11. Displayed Versus Non-Displayed Indexes : 197

Chapter 12. Syntax : 205

Chapter 13. Vocabulary Management : 297

Chapter 14. Surrogation : 365

Chapter 15. Locators : 373

Chapter 16. Surrogate Displays : 383

Chapter 17. Arrangement of Displayed Indexes : 391

Chapter 18. Size of Displayed Indexes : 441

Chapter 19. Search Interface : 447

Chapter 20. Record Format : 495

Chapter 21. Full-Text Display : 521

Chapter 22. Conclusion: Implementation and Evaluation : 537

Glossary : 543

Bibliography : 567

Index : 591

About the Authors : 617

Summary of Contents

Part I. Chapter 1. Introduction and Background Issues : 7

Assumptions and steps required before design begins; terminology; standards and codes of practice, types of databases.

Part II: Design Decisions : 43

Chapter 2. Subject Scope and Domain : 47

Defining the kinds of questions an information retrieval (IR) database will answer, with respect to subject or topical areas and also the operational and cultural domains of potential users; role of subject scope and domain as guide for database producers in the selection of messages and the creation of document descriptions and indexes; role as guide for potential users in the selection and navigation of appropriate IR databases.

Chapter 3. Documentary Scope : 73

Kinds of messages, texts and documents appropriate for fulfilling subject scope and user preferences; definition of searchable documentary features.

Chapter 4. Documentary Domain : 87

Domain (territory) covered in seeking sources for documents and documentary information; methods for discovery, location, and acquisition.

Chapter 5. Display Media : 91

Effective media for the presentation of IR databases.

Chapter 6. Documentary Units : 103

Appropriate size of documentary units to be presented to users; parts of documents versus complete documents and collections of documents.

Chapter 7. Indexable Matter : 111

Portions of documentary units to be analyzed for indexing.

Chapter 8. Analysis and Indexing Methods : 117

Nature of, and appropriate use of, automatic computer-based text analysis and indexing methods versus human intellectual analysis and indexing of messages.

Chapter 9. Exhaustivity : 177

Appropriate level of detail in the indexing of documentary units.

Chapter 10. Specificity : 185

Appropriate level of specificity in indexing vocabulary.

Chapter 11. Displayed Versus Non-Displayed Indexes : 197

Support for multiple search modes: displayed indexes for browsing and visual inspection of headings; non-displayed indexes for machine matching of search statements with message, text, and document representations or full text.

Chapter 12. Syntax : 205

Creating intelligible index headings for browsing and visual inspection; syntactic support for user creation of search statements for computer matching.

Chapter 13. Vocabulary Management : 297

Assistance in dealing with alternative and variant terminology, multiple meanings, and conceptual relations among terms.

Chapter 14. Surrogation : 365

Representing messages, texts, and documentary units within the IR database.

Chapter 15. Locators : 373

Links between representations and full documentary units.

Chapter 16. Surrogate Displays : 383

Options for the display of representations in different contexts.

Chapter 17. Arrangement of Displayed Indexes : 391

Options in the display of browsable indexes: alphanumeric versus classed/relational arrangements.

Chapter 18. Size of Displayed Indexes : 441

Estimating the size of indexes when constraints are present (mainly for printed back-of-the-book indexes).

Chapter 19. Search Interface : 447

Presenting IR databases to users; displaying content and search options.

Chapter 20. Record Format : 495

Size and structure of documentary unit records for supporting design features.

Chapter 21. Full-Text Display : 521

Options for analysis, encoding, and display of full-text documentary units.

Chapter 22. Conclusion: Implementation and Continuing Evaluation : 537

This book is about design. Once the design is complete, it's time to begin testing, implementation, and ongoing evaluation.

Glossary : 543

The glossary contains most of the definitions for terms included in the main text of the book, arranged in alphabetical order. Most of the definitions will refer back to the sections of the book where the concept defined is discussed in more detail, so the glossary can be used as an index to key concepts.

Bibliography : 567

This is the comprehensive bibliography for the book. Each citation concludes with references to the sections where the item is cited, so the bibliography can be used as an index to discussions of cited works.

Index : 591

The alphabetical subject index was compiled using the NEPHIS system, which is described in chapter 12 on syntax, section 12.2.2.3. Index headings refer to sections or paragraphs, not to page numbers.

About the Authors : 617

Full Table of Contents

Prefatory Material : 1

0.1. Foreword, by Jessica Milstead : 1
0.2. Preface, by James D. Anderson : 2
0.3. Acknowledgments, by James D. Anderson : 4
0.4. Special Thanks to Scholars and Practitioners of IR for the Use of Their Work : 4
0.5. Bibliographic Citations : 5
0.6. Dedication : 5

Part I. Chapter 1. Introduction and Background Issues : 7

1.1. Purpose : 7
1.2. Assumptions : 9
1.3. Terminology : 10
1.4. Standards and Codes of Practice : 24
1.5. Types of IR Databases : 30
1.5.1. Kinds of Objects Represented in Index Terms, Headings, and Entries : 31
1.5.2. Kinds of Index Terms Used : 32

1.5.3. Kinds of Indexable Matter Used : 33
1.5.4. Presentation and Methods for Searching : 33
1.5.5. Arrangement of Index Entries : 34
1.5.6. Methods for Analysis : 35
1.5.7. Methods for Term Selection in Indexing : 35
1.5.8. Methods for Term Combination in Searching : 35
1.5.9. Kinds of Documents Being Indexed : 36
1.5.10. Media of IR Databases : 37
1.5.11. Proximity of Documents Being Indexed : 37
1.5.12. Size of Documentary Units : 38
1.5.13. Periodicity of IR Databases : 38
1.5.14. Authorship of IR Databases : 39
1.5.15. Continuing Examples : 39
1.6. IR Databases Versus Other Types of Databases: A Recap : 39
1.6.1. Two Types of Databases : 40
1.6.2. IR Databases : 40

Part II: Design Decisions : 43

Chapter 2. Subject Scope and Domain : 47

2.1. Specialized Categories : 57
2.2. Presenting the Subject Scope and Domain to Users : 61
2.3. Ranganathan's Facets : 63
2.4. Why Bother? : 66
2.5. Our Examples : 67
2.5.1. A Book Index : 67
2.5.2. An Indexing and Abstracting Service : 69
2.5.3. A Full-Text Encyclopedia/Digital Library : 72

Chapter 3. Documentary Scope : 73

3.1. Authorship : 75
3.2. Titles : 76
3.3. Media : 76
3.4. Forms, Formats, Genres : 77
3.5. Periodicity : 80
3.6. Intended Audience or Level : 81
3.7. Methodological Approaches, Points of View, Biases, Kinds of Treatment : 81
3.8. Language : 81
3.9. Place of Creation, Manufacture, or Publication : 82
3.10. Time of Creation, Manufacture, or Publication : 82
3.11. Specific Documents : 82
3.12. Qualitative Criteria : 82
3.13. Features Versus Topics : 83
3.14. Our Examples : 83
3.14.1. A Book Index : 83
3.14.2. An Indexing and Abstracting Service : 84

3.14.3. A Full-Text Encyclopedia/Digital Library : 86

Chapter 4. Documentary Domain : 87

4.1. Primary Versus Secondary Sources : 88
4.2. Monitoring and Covering Documentary Domain : 88
4.3. Our Examples : 89
4.3.1. A Book Index : 89
4.3.2. An Indexing and Abstracting Service : 89
4.3.3. A Full-Text Encyclopedia/Digital Library : 90

Chapter 5. Display Media : 91

5.1. Paper : 91
5.1.1. Card Files : 92
5.1.2. Books : 92
5.1.3. Optical Coincidence (Peek-a-Boo) Retrieval on Cards : 94
5.2. Microforms : 98
5.3. Electronic Media : 98
5.3.1. Online Databases : 99
5.3.2. CD-ROM and Other Machine-Readable Disks : 99
5.3.3. World-Wide Web : 99
5.4. Codes and Symbols : 100
5.5. Our Examples : 100
5.5.1. A Book Index : 100
5.5.2. An Indexing and Abstracting Service : 101
5.5.3. A Full-Text Encyclopedia/Digital Library : 101

Chapter 6. Documentary Units : 103

6.1. Some Examples : 104
6.2. One of the Few “Laws” in Information Retrieval : 106
6.3. Documentary Units Versus Surrogates : 107
6.4. Multiple Documentary Units : 108
6.5. Our Examples : 108
6.5.1. A Book Index : 108
6.5.2. An Indexing and Abstracting Service : 109
6.5.3. A Full-Text Encyclopedia/Digital Library : 109

Chapter 7. Indexable Matter : 111

7.1. Typical Examples of Indexable Matter : 112
7.1.1. Titles : 112
7.1.2. Titles and Abstracts : 112
7.1.3. Preliminary Matter : 113
7.1.4. Initial Paragraphs : 113
7.1.5. Internal Indexes : 113
7.1.6. Reference Citations : 113
7.1.7. Opening Screens of Web Sites : 114
7.1.8. Full Texts : 114
7.1.9. Types of Messages : 114

7.2. Indexable Matter Versus Subject Scope : 114
7.3. Accuracy of Indexing : 115
7.4. Our Examples : 115
7.4.1. A Book Index : 115
7.4.2. An Indexing and Abstracting Service : 115
7.4.3. A Full-Text Encyclopedia/Digital Library : 116

Chapter 8. Analysis and Indexing Methods : 117

8.1. Research Comparing Automatic and Human Indexing : 119
8.2. Human Analysis and Indexing : 123
8.2.1. Cognition Versus Social Construction in Human Analysis and Indexing : 130
8.2.2. Human Indexing Rules : 135
8.2.2.1. Human Indexing Rules for Image Text : 141
8.2.2.2. Human Indexing Rules Based on Probabilistic Analysis : 142
8.3. Automatic Analysis and Indexing : 145
8.3.1. In the Beginning Was the Word : 146
8.3.2. Simple Keyword Indexing : 148
8.3.3. Negative Vocabulary Control: Stop Lists : 149
8.3.4. Counting Words : 149
8.3.5. Comparative Counting and Weighting : 149
8.3.6. Improving the Count: Stemming : 150
8.3.7. Natural Word Distributions : 151
8.3.8. Words Versus Phrases : 156
8.3.9. Managing Vocabulary in Automatic Indexing : 159
8.3.10. Automatic Vocabulary Management : 161
8.3.11. Clustering : 164
8.3.11.1. Latent Semantic Indexing : 166
8.3.12. Citation Indexes : 167
8.3.12.1. Bibliographic Coupling : 167
8.3.12.2. Co-Citation : 168
8.3.13. Relevance Feedback : 168
8.4. Subject Analysis and Indexing in Indexing and Abstracting Services : 169
8.5. Growing Role of Automatic Analysis and Indexing : 171
8.5.1. Censorship or Guidance? : 173
8.6. Our Examples : 175
8.6.1. A Book Index : 175
8.6.2. An Indexing and Abstracting Service : 175
8.6.3. A Full-Text Encyclopedia/Digital Library : 176

Chapter 9. Exhaustivity : 177

9.1. Recall and Precision : 178
9.1.1. A Definition of Recall : 178
9.1.2. A Definition of Precision : 179
9.1.3. The Impact of Exhaustivity on Recall and Precision : 180
9.2. The Calculation of Exhaustivity: Terms Versus Headings : 180
9.3. Our Examples : 181
9.3.1. A Book Index : 181

9.3.2. An Indexing and Abstracting Service : 182
9.3.3. A Full-Text Encyclopedia/Digital Library : 182

Chapter 10. Specificity : 185

10.1. Definitions of Specificity : 185
10.2. Relations between Exhaustivity and Specificity : 190
10.3. Examples of Specificity : 191
10.4. Practical Specificity : 191
10.5. Impact of Specificity on Precision and Recall : 192
10.6. Impact of Specificity on Vocabulary Size : 193
10.7. Specificity Versus Syntax : 194
10.8. Our Examples : 195
10.8.1. A Book Index : 195
10.8.2. An Indexing and Abstracting Service : 195
10.8.3. A Full-Text Encyclopedia/Digital Library : 195

Chapter 11. Displayed and Non-Displayed Indexes : 197

11.1. Displayed Indexes in Electronic Media : 198
11.2. Research on Browsing in Information Retrieval : 201
11.3. Design of Displayed and Non-Displayed Indexes : 202
11.4. Our Examples : 203
11.4.1. A Book Index : 203
11.4.2. An Indexing and Abstracting Service : 203
11.4.3. A Full-Text Encyclopedia/Digital Library : 204

Chapter 12. Syntax : 205

12.1. Precoordinate and Postcoordinate Syntax : 206
12.2. Precoordinate Syntax for Displayed Indexes : 208
12.2.1. Subject Heading Syntax: *Library of Congress Subject Headings* (LCSH) : 208
12.2.1.1. *Medical Subject Headings* (MeSH) : 227
12.2.1.2. Principles for Subject Heading Systems : 228
12.2.2. String Syntax : 231
12.2.2.1. Rotated Term Syntax : 231
12.2.2.2. Faceted Syntax (PRECIS, CIFT) : 234
12.2.2.2.1. Converting LCSH to Faceted Syntax : 241
12.2.2.3. Ad Hoc String Syntax (NEPHIS) : 243
12.2.3. Relational Syntax : 247
12.2.3.1. Syntagmatic Relationships : 250
12.2.4. Classification Syntax : 251
12.2.4.1. Chain Syntax : 257
12.2.5. Natural Language Syntax : 258
12.2.5.1. KWIC Syntax : 260
12.2.5.2. KWOC Syntax : 261
12.2.5.3. KWAC Syntax : 261
12.2.6. Permuted Syntax : 262
12.2.7. Ad Hoc Syntax : 263
12.2.7.1. Combining Ad Hoc Syntax with Systematic Syntax : 266

12.2.8. Syntactic Cross References : 269
12.3. Postcoordinate Syntax for Non-Displayed Indexes : 269
12.3.1. Exact Match (Boolean) Syntax : 273
12.3.2. Best Match (Weighted Term) Syntax : 275
12.4. Our Examples : 280
12.4.1. A Book Index : 280
12.4.2. An Indexing and Abstracting Service : 284
12.4.3. A Full-Text Encyclopedia/Digital Library : 294

Chapter 13. Vocabulary Management : 297

13.1. The Vocabulary Problem : 297
13.2. Research on Vocabulary Issues : 300
13.3. Vocabulary Solutions : 301
13.3.1. Syndetic Structure in Displayed Alphabetical Indexes : 306
13.3.2. Indexing Thesauri : 313
13.3.2.1 Examples of Indexing Thesauri : 318
13.3.3. End-user Thesauri : 332
13.3.3.1. Compiling an End-User Thesaurus : 333
13.3.3.1.1. Sources of Terms : 333
13.3.3.1.2. Selecting Terms : 333
13.3.3.1.3. Categorizing Terms : 337
13.3.3.1.4. Bound Terms Versus Elemental Descriptors : 340
13.3.3.1.5. Term Relationships : 342
13.3.3.1.6. Variant Forms and Equivalent Terms : 352
13.3.3.1.7. Homographs : 353
13.3.3.1.8. Thesaurus Displays : 354
13.3.4. Co-occurrence Term Clustering : 354
13.3.5. Ontologies : 355
13.4. Our Examples : 358
13.4.1. A Book Index : 358
13.4.2. An Indexing and Abstracting Service : 362
13.4.3. A Full-Text Encyclopedia/Digital Library : 364

Chapter 14. Surrogation : 365

14.1. Purpose of Surrogates : 367
14.2. Guidelines and Standards for Surrogates : 367
14.3. Selected Readings on Abstracts and Abstracting : 369
14.4. Surrogates for Machine Searching : 369
14.5. Our Examples : 370
14.5.1. A Book Index : 370
14.5.2. An Indexing and Abstracting Service : 370
14.5.3. A Full-Text Encyclopedia/Digital Library : 371

Chapter 15. Locators : 373

15.1. Our Examples : 380
15.1.1. A Book Index : 380
15.1.2. An Indexing and Abstracting Service : 381

15.1.3. A Full-Text Encyclopedia/Digital Library : 382

Chapter 16. Surrogate Displays : 383

16.1. Our Examples : 388
16.1.1. A Book Index : 388
16.1.2. An Indexing and Abstracting Service : 389
16.1.3. A Full-Text Encyclopedia/Digital Library : 390

Chapter 17. Arrangement of Displayed Indexes : 391

17.1. Alphanumeric Displays : 392
17.2. Alphanumeric Arrangement in Hypertext Displays : 398
17.3. Relational Classified Displays : 412
17.3.1 Display of Dewey Decimal Classification in Hypertext : 417
17.3.2. Constructing and Displaying a Faceted Classification : 421
17.3.2.1. Display of Faceted Classification in Print Media : 433
17.4. Our Examples : 438
17.4.1. A Book Index : 438
17.4.2. An Indexing and Abstracting Service : 439
17.4.3. A Full-Text Encyclopedia/Digital Library : 440

Chapter 18. Size of Displayed Indexes : 441

18.1. Our Examples : 442
18.1.1. A Book Index : 442
18.1.2. An Indexing and Abstracting Service : 445
18.1.3. A Full-Text Encyclopedia/Digital Library : 446

Chapter 19. Search Interface : 447

19.1. Print on Paper Interfaces : 448
19.2. Electronic Interfaces : 458
19.3. Computer Interface Research: Human Computer Interaction : 460
19.4. Our Examples : 465
19.4.1. A Book Index : 486
19.4.2. An Indexing and Abstracting Service : 490
19.4.3. A Full-Text Encyclopedia/Digital Library : 493

Chapter 20. Record Format : 495

20.1. The MARC Format for Cataloging Data : 495
20.2. Record Format for the *MLA International Bibliography* : 500
20.3. Record Format for ABC-CLIO : 503
20.4. Record Format for a Class IR Database : 505
20.5. The Dublin Core Record Format for Internal Metadata : 511
20.5.1. Dublin Core Qualifiers : 513
20.5.2. Dublin Core Example : 513
20.6 Other Metadata Schemas : 514
20.7. Our Examples : 516
20.7.1. A Book Index : 516
20.7.2. An Indexing and Abstracting Service : 517

20.7.3. A Full-Text Encyclopedia/Digital Library : 519

Chapter 21. Full-Text Display : 521

21.1 Linear Versus Hypertext Formats : 521
21.2 Encoding Schemas for Digital Texts : 524
21.3 Browsing Full Texts Online : 530
21.4. Our Examples : 531
21.4.1. A Book : 531
21.4.2. An Indexing and Abstracting Service : 532
21.4.3. A Full-Text Encyclopedia/Digital Library : 534

Chapter 22. Conclusion: Implementation and Continuing Evaluation : 537

Glossary : 543

Bibliography : 567

Index : 591

About the Authors : 617

Figures

Figure numbers begin with a chapter number, followed by a sequential number. After each figure title, locators refer to the chapter and the paragraph number preceding the figure. Thus Figure 2.1 follows paragraph 74 in chapter 2.

Figure 2.1. Opening screen of *Queer Resources Directory* : 2.74

Figure 2.2. Hypothetical preliminary design for an opening screen for the *MLA international bibliography* : 2.74

Figure 2.3. Hypothetical opening webpage screen for the *MLA international bibliography* : 2.77

Figure 2.4. Hypothetical preliminary opening screen design for an indexing and abstracting service for library and information science : 2.94

Figure 2.5. Hypothetical opening screen for *BLISTER — Bibliography of library & information science & technology: evaluation & research* : 2.94.

Figure 5.1. Photographs of the Jonker peek-a-boo IR system, Rutgers University : 5.19

Figure 8.1. Cooper's odds-payoff indexing chart : 8.119

Figure 8.2. Types of clusters, based on Salton (1975a) : 8.216

Figure 12.1. Index entries from *America: history & life* : 12.129.

Figure 12.2. Index entries from the *MLA international bibliography* : 12.145

Figure 12.3. Index entries from *Psychological abstracts* : 12.264

Figure 12.4. Index entries from *Readers' guide to periodical literature* : 12.264

Figure 12.5. Comparison of key attributes of displayed and non-displayed documentary unit indexes versus simple vocabulary indexes : 12.279

Figures 19.1-19.9. Opening IR database screen designs by students : 19.73

Figure 19.10-11. Display of browsable facets in IR database designs by students : 19.73

Figure 19.12-13. Display of browsable alphanumeric indexes in IR database designs by students : 19.76

Figure 19.14-15. Interface displays for advanced electronic searches in IR database designs by students : 19.79

Figure 19.16-17. Display of results from advanced electronic searches, with vocabulary assistance, in IR database designs by students : 19.79

Figures 19.18-19.19. Display of full surrogates in IR database designs by students : 19.84

Figure 19.20. Hypothetical opening page for an indexing and abstracting service in library and information science : 19.06

Figure 20.1. Record format for the *MLA international bibliography and database* : 20.17

Figure 20.2. Record format for the *American history and life* database (ABC-Clio) : 20.22

0. Prefatory Material

Contents of Prefatory Material

0.1. Foreword, by Jessica Milstead.
0.2. Preface, by James D. Anderson.
0.3. Acknowledgments, by James D. Anderson.
0.4. Special Thanks to Scholars and Practitioners of IR for the Use of Their Work.
0.5. Bibliographic Citations.
0.6. Dedication.

0.1. Foreword, by Jessica Milstead.

Users of information retrieval databases generally have no idea of the complexity of these databases, or of the effort that goes into development of their structure. This is as it should be; the design of the database should not have to be the users' problem.

Once upon a time, in the days when the most sophisticated information retrieval databases were library card catalogs (which many users of this book probably don't remember), the issue of just how much users should be expected to learn in order to satisfy their needs from the database was actually a lively one. Today all that has changed. Willingly or not, designers of information retrieval databases and the access systems that support them have come to recognize that users just want the information they need. The less effort required of the users, the happier they will be, and the more likely they are actually to use the database – and therefore the more likely they are to support its continuation and growth.

There is another side to the question, of course. The effort has to be undertaken somewhere. The easier a given database is to use, the more effort is required behind the scenes to assure that its use will be transparent. And the number of decisions to be made and issues to be resolved in the design of an information retrieval database is certainly not obvious to anyone who has not actually undertaken such a task. It's not just a matter of dumping in some text and citations and turning a search engine loose – at least not if the information is valuable and you desire quality results.

Twenty years ago I wrote a small book (*Subject Access Systems: Alternatives in Design*, published by Academic Press) that discussed some of the major issues in information retrieval database design, but I didn't even try to cover every aspect. Jim Anderson has now undertaken this daunting task, and in the online environment, which barely existed when I wrote my book. We had online databases, but they were hard to use, requiring the skills of trained searchers to exploit them fully.

Prof. Anderson has distilled his decades of experience in teaching and design of information retrieval databases to produce a work that covers every aspect of design. He spent a number of years as the chair of a National Information Standards Organization committee charged with revising the standard for indexes. His experience with that committee has enriched this, his *magnum opus*.

Sound textbooks for information retrieval database design have been very few, and none have been as comprehensive as this. I particularly enjoyed the practice of defining some examples, and following them as example cases in every chapter. With these example cases it becomes possible to see how the principles discussed in the chapter would be applied in an actual database.

Prof. Anderson's work will serve a variety of users. It synthesizes much that is known, while bringing to bear its author's insights on issues. The case studies make the work particularly useful as a textbook, but it will also serve as a refresher for those already in the field, and as a reference for all audiences.

0.2. Preface, by James D. Anderson.

• purpose of this book; definition of IR databases : 1

This book is for our students, and for others who aspire to design the best possible information retrieval (IR) databases for every type of clientele and every type of message, in whatever medium or format. The overall objective is maximum effective retrieval of useful messages for each particular user.

• scope of this book : 2

The scope of this book is determined by the features of modern IR databases. The term "information retrieval database" or "IR database" is used in the broadest sense. Increasingly IR databases are designed for and implemented in digital media, but the design principles addressed in this book apply just as much to all media, including print on paper, microfilm or fiche, or even card catalogs. They certainly apply to modern digital libraries. The basic definition for the term "IR database" as used in this book is any database in any medium designed or created for the purpose of discovering and retrieving messages, texts and documents. Thus, it includes the whole gamut of IR databases presented to users via online connections, the world-wide web, CD-ROMs, or in print on paper: indexing and abstracting services (regardless of medium), library catalogs (including OPACs: online public access catalogs), bibliographies, and indexes, including back-of-the-book indexes (which can now be presented electronically with electronic books!). This and related definitions are expanded in the first part of the book.

• organization of this book; fundamental issues of IR database design : 3

This book is organized around twenty fundamental design issues in IR database design. See the table of contents for a summary of these issues. These are issues that I have identified during twenty-five years of investigation, teaching, design, evaluation, and creation of IR databases. These issues were further refined between 1991 and 1997 by Committee YY of the National Information Standards Organization (NISO). This committee, which I chaired, focused primarily on indexing and indexes, which are fundamental components for all IR databases. But in fact, the committee addressed all of the fundamental design issues for IR databases.

• NISO Committee YY: new standard for indexes : 4

NISO had charged Committee YY with developing a new standard for indexes that would cover every type of index:

- whether produced by human intellectual analysis or machine algorithm;
- whether presented in print or electronic media;
- whether designed for searching by human visual inspection of index headings or by electronic computer matching of search terms against document records or full text; and

● regardless of the code, medium or format of the messages being indexed (linguistic, musical, pictorial, mathematical; visual, aural, tactile; print, microform, film, video, electronic, digital, etc.; monographic or serial; in any genre or format).

● NISO technical report on design of indexes : 5

The results of the work of this committee, based on input from the NISO member organizations most concerned about the design of IR databases as well as hundreds of individual information professionals, was published in a technical report: *Guidelines for indexes and related information retrieval devices* (Anderson 1997a).

● publications related to this book : 6

This NISO technical report covers the scope of this book in a highly compressed format. Another even briefer presentation of the issues covered in this book, but in a less technical and more discursive style, is contained in my encyclopedia article: "Organization of knowledge," in the *International encyclopedia of information and library science* (Anderson 1997b, revised 2002).

● views of Milstead (Jessica L.) on IR database design : 7

This book rests on the foundations of research and practice in information retrieval, IR database design, indexing and cataloging by persons too numerous to mention. Some of their publications are cited in the bibliography at the end of this book. Here I want to acknowledge one book in particular, which has served as a precursor and exemplar for this book: *Subject access systems: alternatives in design* by Jessica L. Milstead (1984). Like this book, Milstead's emphasis and purpose were to foster intelligent design of IR databases. Although organized differently, Milstead covers many of the same points as we do, and we hardly ever disagree! I highly recommend her book. It has aged very well!

● views of Bates (Marcia J.) on IR database design : 8

I also want to pay special tribute to the work of Marcia Bates. One of her earliest articles, "Rigorous systematic bibliography" (1976) got me started in my analysis of IR database design principles. In that article, she focused on attributes of scope and documentary domain, and also on explicit criteria for selectivity. Since then, she has been a leader in research and commentary on essential aspects of the design and use of IR databases. She summarizes much of her work, and that of others, in "Indexing and access for digital libraries and the internet: human, database, and domain factors" (1998). Here she discusses such essential issues as human factors, indexing and searching terminology and structures, statistical distribution properties of documents in collections and databases, and subject domain-oriented indexing. In this book, her work has been especially influential in the chapters on scope and domain, vocabulary management, and interface design.

● views of Chowdhury (Gobinda G.) on information retrieval : 9

Most books on information retrieval focus on particular approaches to this broad field, such as human indexing, library cataloging and classification, or automatic indexing. A recent book that adopts the same kind of broad view as this book is G. G. Chowdhury's *Introduction to modern information retrieval* (1999). The organization and focus of our two books is different, but readers who desire another viewpoint can benefit from Chowdhury's book.

● Pérez-Carballo (José) as co-author: 10

Dr. José Pérez-Carballo, my former colleague in the School of Communication, Information, and Library Studies at Rutgers University, has joined me as co-author of this book. He has helped with the entire book, but he has taken special responsibility for sections on automatic indexing and interface design.

• purpose of this book: 11

I hope that this detailed concentration on the fundamental decision points of IR database design will help members of the information professions to consider all the options, and then to design and create better IR databases. Our society needs the best possible IR databases to cope with the ever growing explosion of information on the internet and the world-wide web, as well as in older print formats, video, film, audio, and electronic formats.

0.3. Acknowledgments, by James D. Anderson.

• acknowledgment to students : 12

Many thanks go to my students who have used and critiqued drafts of this book: Debbie Abrams, Jane Achola, Michael Angeles, Shawn Armington, Robert Barbanell, Frances Berman, Michele Lisoski Bond, Elana Broch, Linda Brown, John Burchard, Dorothea E. Clark, Susan Clark, Lisa Coats, Kathleen Creegan, Thomas M. Dolan, Lisa Ellis, Olga Evanusa-Rowland, Loisann Griglak, Ted E. Hamer, Lonnie Johnson, Mary Kearns-Kaplan, Richard K. Kearney, Michael Knies, Scott Kushner, Mariann E. Lucas, Marygrace Luderitz, Ruth Eleanor Lufkin, Mary Marks, Sal Mazzola, Mary McMahon, Daniel Noonan, Megan Palasciano, Antonio M. Pasqualoni, Beth Patterson, José Fernando Peña, Fran Pfeffer, Frances Pinto, Laura Poll, Jill Ratzan, Robert Rittman, Vivian Thiele, Regan L. Tuerff, Susan Turkel, David Utz, Mary O. Walker, Renee Watson, Karen Wenk, Melissa Yonteck, and Zhu Xuening (Sean). Terry Edwards was an especially careful reader, checking not only text for sense and typos, but also the embedded index strings! Jill Ratzan gave parts of the final text a rigorous perusal.

• 13

I thank them for their valuable editorial assistance. As they worked with earlier drafts of this book, they were very good at pointing out defects.

• acknowledgment to Milstead (Jessica L.), Wellisch (Hans H.), members of NISO Committee YY, and executive director of NISO : 14

Special gratitude goes to my long-term colleagues and primary mentors in the world of indexing, information retrieval, and information science, Dr. Jessica Milstead and Dr. Hans Wellisch, who read intermediate drafts of this book and made many excellent suggestions, most of which I endeavored to implement. I also thank the members of NISO Committee YY, who worked closely with me for many years (1991-1997) on the issues addressed in this book: Barbara Anderson, Knight-Ridder Information, Inc.; Catherine Grissom, U.S. Department of Energy, Office of Scientific and Technical Information; Nancy Mulvany, Bayside Indexing Service; Barbara Preschel, Public Affairs Information Service; Deborah Swain, IBM and Society for Technical Communication; and (again) Hans Wellisch, University of Maryland; and also (again) Jessica Milstead, our liaison from the NISO Standards Development Committee, and Patricia Harris, NISO executive director, both of whom shepherded our work with expertise, care and compassion!

0.4. Special Thanks to Scholars and Practitioners of IR for the Use of Their Work.

• acknowledgment to scholars and practitioners of special importance : 15

This book rests squarely on the work of hundreds of colleagues in the world if IR and IR database design. Every author and every published work is listed in the bibliography, but here we want to give special thanks to those colleagues and organizations whose work we have used most extensively, sometimes with extensive quotes. I trust that our use has been within the bounds of

scholarly "fair use," but beyond the legalities of use and attribution, we want to express our sincere appreciation for and dependence on their work — indeed, these authors and organizations are co-authors of this work with us:

ABC-Clio, American Library Association, Marcia J. Bates, *Bliss Bibliographic Classification,* John P. Comaromi, William S. Cooper, Timothy C. Craven, *Dewey Decimal Classification*, Tamas E. Diszkocs, Karen Markey Drabenstott, Dublin Core, *Eurovoc Thesaurus*, Jason, Farradane, *FOLDOC: The Free On-Line Dictionary of Computing*, Bernd Frohmann, Rebecca Green, Stephan Greene, Donna Harman, David Harper, Marti Hearst, Birger Hjørland, Susan Hockey, Robert R. Korfhage, Library of Congress, Gary Marchionini, Jessica L. Milstead, Modern Language Association of America, National Information Standards Organization, Miranda Pao, A. Steven Pollitt, S. R. Ranganathan, Ronald E. Rice, Rutgers University Libraries, Gerard Salton, Ben Schneiderman, Dagobert Soergel, Karen Sparck Jones, Elaine Svenonius, *Unesco Thesaurus*, Brian C. Vickery, Diane Vizine-Goetz, Bella Hass Weinberg, Hans H. Wellisch, Patrick Wilson.

• acknowledgment to students : 16

We also give special thanks to the students of James D. Anderson who shared their design work to help illustrate concepts in chapter 19:

Matthew Brown, Melissa Hoffman, Eric J. Johnson, Veronica Meyer, Minsoo Park, J. Fernando Peña, Elizabeth Pregill, Robert Rittman, Enola Romano, Lori A. Rowland, Jennifer Schroth.

0.5. Bibliographic Citations.

• style for bibliographic citations : 17

All publications cited in this book are listed in alphabetical order in the bibliography at the end of the book. Citations are presented in accordance with the U.S.A. national standard ANSI/NISO Z39.29-1979. *Bibliographic references* (American National Standards Institute 1979). A revision of this standard was approved in 2003. The only significant change for our purposes was moving the placement of dates for periodical articles from the end of the citation, after volume and issue numbering and pagination, to prior to the volume and issue numbering (National Information Standards Organization 2004?). We did not adopt this small change.

0.6 Dedication.

We dedicate this work to Rafael and Dwayne, the wind beneath our wings.