Information Retrieval Design. James D. Anderson and Jose Perez-Carballo. East Brunswick, NJ: Ometeca Institute, Inc., 2005. 617 pp. $35.00. (0-9763547-0-5).

Information Retrieval Design is a textbook that aims to foster the intelligent user-centered design of databases for Information Retrieval (IR). The book outlines a comprehensive set of twenty factors, chosen based on prior research and the authors' experiences, that need to be considered during the design process. The authors provide designers with information on those factors to help optimize decision-making. The book does not cover user needs assessment, implementation of IR databases or retrieval systems, testing or evaluation.

Most textbooks in IR do not offer a substantive walkthrough of the design factors that need to be considered when developing IR databases. Instead they focus on issues such as the implementation of data structures, the explanation of search algorithms and the role of human-machine interaction in the search process. The book touches on all three, but its focus is on designing databases that can be searched effectively, not the tools to search them. This is an important distinction, despite its title this book does not describe how to build retrieval systems. Professor Anderson utilizes his wealth of experience in cataloguing and classification to bring a unique perspective on IR database design that may be useful for novices, for developers seeking to make sense of the design process and for students as a text to supplement classroom tuition.

The Foreword and Preface (by Jessica Milstead and James Anderson respectively) are engaging and worthwhile reading. It is astounding that it has taken some 20 years for anyone to continue the work of Milstead and write as extensively as Anderson does about such an important issue as IR database design. The remainder of the book is divided into two parts: *Introduction and Background Issues* and *Design Decisions*.

Part One is a reasonable introduction and includes a glossary of the terminology that authors use in the book. It is very helpful to have these definitions early on, but the subject descriptors in the right margin are distracting and do not serve their purpose as access points to the text. The terminology is useful to have as the authors definitions of concepts do not fit exactly with what is traditionally accepted in IR. For example, they use the term "message" to refer to what would normally be called "document" or "information object", and do not do a good job at distinguishing between "messages" and "documentary units".

Part Two describes components and attributes of IR databases to help designers make design choices. The book provides them with information about the potential ramifications of their decisions and advocates a user-oriented approach to making them. Chapters are arranged in a seemingly sensible order based around these factors and the authors remind us of the importance of integrating them. The authors are skilled at selecting the important factors in the development of seemingly complex entities, such as IR databases. However, the integration of these factors, or the interaction between them, is not handled as well as perhaps should be. Factors are presented in the order in which

authors feel they should be addressed but there is no chapter describing how the factors interact. The authors miss an opportunity at the beginning of Part Two where they could illustrate using a figure the interactions between the twenty factors they list in a way that is not possible with the linear structure of the book.

What follows is a subject-based review. To aid the flow of the review and reduce redundancy, chapters are grouped together by subject.

Chapters Two to Five: Scopes, Domains and Display Media (pp. 47 – 102). Chapter Two describes the role and the value of the scope (i.e., the kinds of topical questions an IR system must respond to) and the domain (i.e., the subject set in the context of the work or life situation) within which IR databases must operate, and differences between them. In Chapter Three the authors describe the type of documents that users can retrieve using a database in terms of non-topical features (e.g., authorship, media, intended audience). The authors suggest that document scope can be useful when the search is constrained by document types (e.g., format, medium and periodicity), not the topical content of the document. Chapter Four describes the document domain in which the IR database must operate, the difference between primary and secondary sources and how this document domain can be monitored and covered. In Chapter Five the authors describe the media that can be used to present information to end-users. The authors comprehensively list different forms of media that can be used in IR databases.

Chapters Six to Eight: Documents, Analysis and Indexing (pp. 103 – 176). Chapter Six debates the issue of the optimal size of documents for retrieval from IR databases. The authors stress the importance of selecting the appropriate document type for retrieval. In Chapter Seven the role of the information indexed by the IR system is described and a comprehensive overview of the types of information indexed is provided. In Chapter Eight the authors describe how language text is indexed (either by humans or automatically), although does not focus on the indexing of multimedia or other non-language texts. The authors spend much time on human indexing and supplement their views with those of others. This is the first chapter in the book where traditional IR techniques such as stemming, stopword removal and term frequencies are discussed.

Chapters Nine to Ten: Exhaustivity and Specificity (pp. 177 – 196). In Chapter Nine the authors discuss "exhaustivity" (i.e., how well the topics and features of documents are described). This is the first time traditional measures of IR effectiveness such as recall and precision are mentioned. In Chapter Ten the authors outline the role of "specificity" (i.e., the semantic term-document relation between the index terms and the topics discussed in documents or the features of documents). The authors highlight the ambiguity associated with the concept and describe other researchers' views of it

Chapters Eleven to Thirteen: Displayed/Non-displayed Indexes, Syntax and Vocabulary Management (pp. 197 – 364). Although not highlighted explicitly by the authors, there is a marked shift between Chapters Ten and Eleven, and the role of the user in the retrieval process become more obvious. In Chapter Eleven the authors examines

the role of displayed, browsable indexes for human inspection. This chapter leads directly into the discussion of Syntax in Chapter Twelve and provides a reasonable (if short) treatment of indexing display in IR and a summary of browsing in IR. Chapter Twelve provides an extremely thorough discussion of the role of syntax as patterns for putting terms together for indexing or searching. Chapter Thirteen describes the management of vocabulary in IR, including the problems associated with differences in vocabulary and the use of alphabetic indexes, thesauri and ontologies. The categorization of searches offered in this chapter could be beneficial to readers, but perhaps the authors could have given more attention to related work in this area.

Chapters Fourteen to Sixteen: Surrogation, Locators and Surrogate Displays (pp. 365 – 390). In Chapter Fourteen the content and uses of surrogates are defined (including guidelines and standards for surrogates). Chapter Fifteen addresses locators, an essential part of surrogates, and Chapter Sixteen describes the progressive presentation of surrogate content from surrogate information to document full-text.

Chapters Seventeen to Eighteen: Arrangement and Size of Displayed Indexes (pp. 391 – 446). In Chapter Seventeen the book describes how indexes should be arranged for presentation to users. Chapter Eighteen discusses the size of these displayed indexes. It seemed unnecessary to have a separate chapter for each of these closely-related issues.

Chapters Nineteen to Twenty-One: Search Interface, Record Format and Full-Text Display (pp. 447 – 536). The authors begin Chapter Nineteen by stressing that the

most important element in IR databases is the user interface. This is rather surprising bearing in mind the amount of time the authors devote in the book up until this point describing the role of components hidden from the user's view. The review of research in this area is sufficient for the purpose of the book, but I would have expected a more detailed account of the role of Human-Computer Interaction in IR. Chapter Twenty describes the record format that defines how the data will be tagged, labelled and stored in the database. Chapter Twenty-One addresses the format of documents and how the full-text should be presented to users.

Chapter Twenty-Two: Implementation and Evaluation (pp.537-541). The book concludes with a brief discussion of issues in implementation, testing and evaluation. Examples aside, this is the only time that these fundamental issues are addressed in any way in the book. The chapter is a review of currently available commercial technologies (e.g., EndNote, Reference Manager) rather than a description of how IR systems can be developed. Since the book tries to make sense of the design space it may have been useful to point readers to further reading that operationalizes many of the issues raised. Testing and evaluation are covered briefly and organized by evaluation topic. The book lacks a conclusion or takeaway message and would benefit from the inclusion of both.

The book uses examples to illustrate the concepts it describes. These include examples of a book index, an indexing and abstracting service and new encyclopaedia/digital library. These are used at the end of each chapter to reinforce the concepts introduced therein and readers can follow the same three examples throughout

the book. These are useful for readers but also shows that the authors understand who their readers are likely to be and the types of questions such readers would ask. The "book index" example is somewhat trivial and is perhaps only useful to novices but the others may be helpful to all readers. Surprisingly, aside from a few cursory references, the book almost completely ignores the World Wide Web and its current role in information dissemination. It may have been useful to include more references to this domain, as many of the ideas discussed have been implemented there in some form.

The index of the book is comprehensive and the "related terms to consider" are helpful and should have perhaps have been included for more than only a few terms. The inclusion of section numbers after each citation in the bibliography is helpful in pinpointing potentially useful parts of the book.

Those interested in using the book to design IR databases can work through the chapters in the order provided and end up with a set of requirements for database design. The steps outlined in this book can be rearranged in numerous orders depending on the particular circumstances. This book would benefit from a discussion of what orders are appropriate for different circumstances and how the requirements outlined interact.

I come away from *Information Retrieval Design* with mixed, although mainly positive, feelings. Even though the aims of this book are made clear from the outset it was still a disappointment to see issues such as implementation and evaluation covered in only a cursory manner. The book is very well structured, well written and operates in a

part of the space that has been neglected for too long. The authors whet my appetite with discussion of design and I would have liked to have heard a bit more about what happens in requirements elicitation before the design issues have been identified and in implementation after they have been addressed.

Overall, the book is a comprehensive review of previous research supplemented by the authors' views on IR design. This book focuses on breadth of coverage rather than depth of coverage and is therefore potentially of more use to novices in the field. The writing style is clear and the authors' knowledge of the subject area is undoubted. I would recommend this book to anyone who wants to learn about IR database design and take advantage of the experience and insights of Anderson, one of the visionaries in the field.