Indexing, teaching of
See: Information retrieval design

James D. Anderson

Places the teaching of (about) indexing into the context of the 20 fundamental attributes of all information retrieval databases or systems, ranging from back-of-the-book indexes to physical and digital libraries, indexing and abstracting services, and online or web-based message/document retrieval services.

> I do not believe that indexing can be taught. Rules, and the reasons for following or not following them, can be presented. Various index formats can be discussed. However, the ability to objectively and accurately analyze text and to produce a conceptual map that directs readers to specific portions of the text involves a way of thinking that can only be guided and encouraged, not taught. . . . Indexing cannot be reduced to a set of steps that can be followed. (Nancy Mulvany, *Indexing Books,* 1994, pp. vii–viii)

Introduction

I agree with Nancy Mulvany that indexing *per se* cannot be taught, so in my teaching I try to present everything I can *about* indexing, with the hope of enabling my students to make the best choices about how to approach an indexing task.

I take a very broad view of indexing, encompassing everything we can to do to point to ('to indicate') messages of interest. I include all types of messages of every size (from a single sentences to entire libraries; from a postage stamp or a segment of a photograph to an entire film, video or archive). Messages come in an enormous variety of formats, media and texts. Texts are not at all limited to language texts; I subscribe to the semiotic view of 'text' as a meaningful collection of symbols, so my texts include visual images (photographs, paintings, sculpture, architecture), sound images (music), and performance images (theater, dance, etc.).

Over the years as an indexer, librarian and designer of indexing and information retrieval (IR) systems, I have settled on a list of features that I see in every index – in every IR system. I contend that every index – every IR database or IR system – represents decisions about these 20 features. The term 'IR database' stands for every variety of index, bibliography, digital or physical library – in other words, any collection (database) designed for the retrieval of messages by their topic, content or meaning.

The list of 20 essential features is as follows:

1. Subject scope and domain (specifying the kinds of subject-oriented questions or needs to be addressed, based on a prior needs assessment of potential users and their operational and/or cultural domains).
2. Documentary scope (delimiting kinds and qualities of messages, texts, media, formats, etc. appropriate for addressing the subject scope and domain; identification of searchable features).
3. Documentary domain (describing sources of documents; territory covered in pursuit of useful documents).
4. Display media (for the index or IR system).
5. Documentary units (the units of text analyzed for retrieval).
6. Indexable matter (the portion of text used to represent the documentary unit).
7. Content analysis methods (human intellectual versus machine algorithmic methods, or both).
8. Exhaustivity of indexing.
9. Specificity of indexing terms.
10. Displayed versus non-displayed indexes (options for browsing versus electronic searching).
11. Syntax for index headings in displayed indexes and for search statements to be matched against non-displayed indexes.
12. Vocabulary tracking, management, facilitation.
13. Surrogation for messages and texts.
14. Locators and links among surrogates and documents.
15. Surrogate displays (ranging from brief to full and the arrangement of elements).
16. Arrangement of browsable displayed indexes (alphabetical, relational, hybrid).
17. Size of displayed indexes (in print media).
18. Search interfaces; information visualization; explanation.
19. Record structure; organization of metadata.
20. Display of full text.

My teaching of indexing comprises the examination of these 20 features. Each student analyzes an existing IR database – anything from a back-of-the-book index to large indexing and abstracting (A&I) services presented in print or in electronic hypertext online or via the world wide web. They seek to discover the decisions made for the 20 design features. At the same time, students design their own IR databases – again ranging from back-of-the-book indexes to large A&I services.

This brief description of 'how I teach indexing' guides the reader through the 20 features. I tell my students that these 20 options are like 20 steps, 20 decisions in a recipe for shaping the best possible index or IR system for a particular clientele or application.

In the brief compass of this article, it is not possible go into the detail that is covered in the classes. Here, I do little more than sketch in some of the issues.

1. Subject scope and domain

The subject scope and domain serves several functions. It reflects the designer's analysis of the needs, interests and desires of users, summarizing the kinds of topics they are likely to seek. It guides the collection of appropriate documents for A&I services and digital libraries. And it reminds human indexers of what to look for in texts. Even in a back-of-the-book index, it is good to decide beforehand whether certain categories of topics will be indexed. Will the index, for example, have entries for cited authors, such as: 'Mulvany, Nancy – views on teaching of indexing'?

The subject domain refers to the work or life domains of potential users – the contexts out of which their information needs come. The same topics would be treated very differently within the domains of researchers or practitioners or patients or consumers; if patients or consumers, are they adults or children? How the subject scope is presented should be influenced by the work and life domains of IR users (Hjørland, 1997).

The subject scope is, in effect, an indexing of the entire IR database. In order to encompass the often broad subject scopes of IR databases in approximately 10–20 categories, I encourage students to do a facet analysis of the subject scope, beginning with Ranganathan's basic PMEST categories: Personality (or entities, concrete and abstract, including living beings, groups, institutions, texts, natural objects, and artifacts); Material (including attributes, properties, and also parts of entities); Energy (including actions, operations, processes and events); Space (including places and environments); and Time (Ranganathan, 1965).

In the electronic environment in which we all work today, the subject scope provides the framework for a faceted classification of the content of the IR database, which can be used to create browsable relational classified displays – indispensable for giving users an overview of the content of the database and access to broad categories of information contained within its documents.

2. Documentary scope

Documentary scope simply describes the kinds of documents included in the IR database, and the document features that should be indexed for retrieval, such as author, title, language, format, date of creation or publication, etc. It also includes criteria for inclusion in collections of documents (A&I services), whether they are to be comprehensive within their subject scopes, or selective. And if they are selective, on what basis? These attributes are stressed in one of my all-time favorite papers by Marcia Bates (1976), in which she builds on the pioneering analysis of Patrick Wilson (1968).

3. Documentary domain

Where and how documents are sought to ensure maximum coverage can be skipped for a back-of-the-book index or an already existing collection of documents, but it is an essential attribute for ongoing A&I services. Two A&I services could have identical subject and document scopes, but the documentary domain of one could be limited to one particular library (in which case it should be called a catalog), while in the other, the documentary domain may cover the entire world, combed by a network of advisors, contacts, and bibliographers. This is another attribute stressed by Marcia Bates (1976).

4. Display media (for the index or IR system)

This is the medium for the IR database itself, in contrast to the media of the documents included. All my students design IR databases for electronic, hypertext media, but they may also design a print-on-paper version. Older, traditional media, such as cards and microforms, have largely been discarded.

5. Documentary units

Documentary units (also called units of analysis) are documents or segments of documents analyzed by humans or by machine algorithms in order to describe their content, topics and/or meaning. These units are typically different for machine and human indexing. For full-text analysis by machine, the units should be small, such as a paragraph of language text, so that the search (indexing) algorithm will take the user to the paragraph with the best match with the search statement.

Human analysis is usually focused on larger units, such as complete periodical articles or even complete books, but of course humans can also pinpoint and describe important illustrations or tables and other special parts of larger documents.

The traditional documentary unit for back-of-the-book indexers has been the page, but that is changing as more and more books are published in digital media. The problem with the page is that it is an artifact of the print-on-paper book format, not of the text (with rare exceptions for certain art books). Thus, when the text moves to digital media, the pages are lost, unless artificially imposed. This makes an index whose headings are linked to pages useless!

The NISO *Guidelines for indexes and other information retrieval devices* (Anderson, 1997) encourage book designers and indexers to use documentary units that are inherent features of texts, such as paragraphs. This means that both designers and indexers need to devise discrete ways of numbering paragraphs so that users can move directly from an index to the relevant paragraphs. This will be an advantage for users who in the past may have had to wade through a dense page to find the paragraph that relates to an index heading of interest.

6. Indexable matter

Indexable matter (or 'analysis base') is the actual portion of text that is analyzed for indexing a documentary unit. This applies both to human and machine indexing. Examples include A&I services that represent (index) entire period-

ical articles by using only abstracts. The Institute for Scientific Information bases its 'Permuterm' indexing only on titles of documents, and its citation indexing only on reference citations. Thus, their indexable matter is titles in the first case, and reference citations in the second.

Much more detailed indexing, as for back-of-the-book indexes, is almost always done on the basis of the full text of documentary units, but I admit that when I am only allowed space for a very skimpy index, I may not look much past topic sentences! In that case, if the paragraph is the documentary unit, the first sentence would be the indexable matter!

7. Analysis and indexing methods (human intellectual; machine algorithm)

Here students choose human intellectual methods and machine algorithms for text analysis. Machine methods are currently almost completely limited to language texts, work on visual texts still being very rudimentary. Students examine such options as simple word indexing (the state of the art in many A&I services, which permit only Boolean searching), or more advanced techniques that allow weighted ranked retrieval, phrase identification, or clustering. As examples of clustering, we look at both latent semantic indexing, where documentary units are identified and retrieved through their association with clusters based on word co-occurrence, and also bibliographic coupling and co-citation, in which documentary units are clustered on the basis of their own shared reference citations or on the basis of being cited frequently together by other writers. Phrase identification techniques have been used to great advantage to provide browsable interfaces for digital libraries (Gutwin et al., 1999).

Methods of human analysis are much harder to pin down, though some work has been done in trying to identify them (Anderson and Pérez-Carballo, 2001). The most promising attempt to guide human analysis, in my view, is providing indexers with a subject scope statement (see above, section 1), which should describe in some detail what indexers should look for. Both the ISO (International Organization for Standardization, 1985) and the BSI (British Standards Institution, 1984) have attempted to do this. Cooper (1978) has also attempted to codify the general rule of thumb that indexers should use a term if users using the same term in a search would want the documentary unit in question. He has created a kind of formula to calculate this situation based on the odds against user satisfaction versus the amount of money a user would be willing to pay for the documentary unit. No one has pursued his suggestions, to my knowledge.

8. Exhaustivity of indexing (analysis and representation)

The detail of indexing is a very important policy consideration for any index, with strong implications for search precision and recall. The simple definition of exhaustivity is how many single concept terms will be used to describe the topics, content or meaning of a documentary unit – as few as only one (much too skimpy in most cases) to more than 100. Here we are *not* concerned with how these terms will be put together or displayed in headings (an issue of syntax), but only with the number of terms, representing the detail of indexing description.

Students generally follow indexing tradition by using a relatively high threshold of importance for human term selection or assignment (low exhaustivity), which should contribute to higher precision, with a possible reduction of recall, while relying on automatic indexing for much more exhaustive indexing, which may lead to higher recall, but at a cost of precision.

9. Specificity of indexing terms

Specificity tends to offset or modify the impact of exhaustivity. It refers to the tightness of fit between the meaning of a term and the topic of a discussion or illustration in a text. Thus, if the text illustrates or discusses a Labrador retriever, but the indexing term is 'dogs', or 'canines', or 'animals', specificity is low and the generic level is increasingly high.

Highly specific indexing terms favor high-precision retrieval. If I use the indexing term 'Labrador retrievers', I should get documentary units that discuss or illustrate Labrador retrievers. However, if I am forced to use the indexing term 'dogs', then of course I will retrieve many items that don't mention Labrador retrievers. On the other hand, recall may suffer a little if there are documentary units on many breeds of dogs, so the indexer uses the term 'dogs' rather than separate descriptors for each breed. The searcher who wants *everything* on Labrador retrievers may miss these documentary units, unless the vocabulary management component links 'Labrador retrievers' to 'dogs' with a note like: 'for comprehensive searches, you may also want to use the broader term DOGS'.

It is an error to equate specificity with number of postings in an IR database. There is no necessary relationship. 'Toxicity' might be very specific to many documentary units in the Toxline database, yet also have a very high number of postings. Similarly, specificity has nothing to do with the breadth or generality of meaning of a term. If disease or diseases in general are under discussion, then 'disease' as a condition or 'diseases' as a class is highly specific to the documentary unit.

Specificity relates to the *relationship* between an indexing term and a topic or feature of a documentary unit. In recent research purporting to study specificity, researchers have tended to conflate the concept with numbers of postings or level of generality, for example in a thesaurus hierarchy. These are related concepts, but they are *not* specificity, and these related concepts *do not* have the impact on retrieval that specificity has.

10. Displayed versus non-displayed indexes (options for browsing vs electronic searching)

In print media, indexes must be displayed for human inspection and browsing. In many electronic IR databases, users never get to see an index – browsing is not possible during

the search process, but only after documentary units are retrieved in response to an electronic term-matching search. (The display of descriptors, as in a thesaurus, does *not* constitute a displayed index to documentary units, which would have multi-term headings representing the content of these documentary units.)

Recent research has emphasized and verified the importance of browsing in the context of electronic IR databases. Electronic term-matching searching is satisfactory for users who know exactly what they want and what to call it, in the indexing language of the IR database, but there are many people who are 'just looking', or who are unsure what they want or what to call it. These are users with 'anomalous states of knowledge', described by my Rutgers colleague Nick Belkin et al. (1982a, b). It is much easier for people in this situation to recognize a good term or heading than to think it up in an IR vacuum!

Recent evidence has also indicated that a growing number of Web 'search engines' and other IR databases are providing browsable displays. Weise (2000) stated that

> . . . for once in the information revolution, the humans are pulling ahead. . . . The computer-automated indexes that powered a majority of the Web's search engines gave ground to Web directories – listings that depended instead on the power of thousands of human minds to harness the limitless information of the Net. . . . In December [1999], the top five search sites on the Net – Yahoo!, AOL, MSN, Netscape and Lycos – were all based mainly on human-generated directories rather than computer-created indexes, according to figures from the market trackers Nielsen/NetRatings.

About the same time, Christine Borgman (2001), a leading information scientist, said pretty much the same thing: 'In the networked world, browsing has supplanted direct searching as the primary means to locate information.'

Consequently, I require all my students to design browsable displayed indexes for their electronic IR databases, including electronic back-of-the-book indexes, and I require that they provide not only alphanumeric displays, but also relational classified displays (see below, section 16).

11. Syntax for representation in displayed indexes or search statements

Syntax provides the means for combining terms in electronic searches and in browsable index headings. For electronic searches, syntactic methods can be grouped into basic Boolean versus weighted term methods, including vector space and probabilistic approaches. All electronic syntaxes often include such additional features as proximity specification, stemming and truncation.

Syntax for browsable headings includes traditional subject headings and the traditional 'ad hoc' headings created, heading by heading, by back-of-the-book indexers. I require my students to opt for more modern, labor-saving syntaxes, often called 'string indexing' methods (Craven, 1986), in which the human indexer provides the terms, and computer algorithms create headings, under all important access points. Craven's NEPHIS (Nested Phrase Indexing System) works particularly well for back-of-the-book indexing, and I always use it when I create such indexes.

12. Vocabulary tracking, management, facilitation

All IR databases must have a vocabulary management component for linking synonymous, equivalent and associated terms (broader, narrower and related). The only question is whether the IR database will provide this assistance to users, or whether users must provide it themselves. I require my students to integrate vocabulary assistance into their search interfaces, whether for electronic term-matching searches or alphanumeric or relational classified browsable indexes. To support such vocabulary management or facilitation, they design thesauri or ontologies that are linked to their IR database search interfaces. Clustering of terms based on co-occurrence can also be a helpful technique for suggesting related terms for electronic searching.

13. Surrogation for messages and texts

Even full-text IR databases need briefer surrogates, so that users have the chance to get overviews of multiple documentary units of potential interest. Here students decide what should be included in their full surrogates, such as citations, abstracts, descriptors, summaries, thumbnail illustrations, and so on.

14. Locators and links among surrogates and documents

Every IR database in whatever medium must have some system for linking surrogates, including index headings (which are usually the briefest surrogates), to larger surrogates and/or full texts. Locators will generally be hypertext links in electronic IR databases. In print databases, the number of locators per heading indicates the number of postings. With hypertext links, the number of postings should always be included. This is very helpful information for a user who must decide whether to pursue a link or locator.

15. Surrogate displays

Surrogates should be displayed in stages, geared to the stages of a search. My students design a range of surrogate displays, ranging from very brief (one line only, such as a full index heading string), to brief, to full. They also determine the arrangement of elements, making sure that they match the search. In a subject search, the surrogate *should not* begin with non-topical information such as authors or titles, but with descriptors or index headings!

16. Arrangement for browsable displayed indexes (alphabetical, relational, hybrid)

Here we deal with a scandal of the Library and Information Science profession – its inability to specify a commonly accepted method for arranging alphanumeric displays. It is likely that the wide disparities in alphanumeric arrangement thoroughly confuse most users, who expect alphanumeric

order to be simple and straightforward. The constraints of one-dimensional displays in print media are contrasted with multi-layered, multi-dimensional hypertext displays in electronic media. (It is another scandal that so many online catalogs still display subject headings as if they were on cards in a card catalog!)

In addition to alphanumeric displays, students design facet-based relational classifications for browsing, so that users can choose facets that match their primary interests and can search several such facets simultaneously. This feature was included in the 1980 design for the MLA International Bibliography database, but was never implemented by vendors (Anderson, 1979, 1980). Now Steven Pollitt and his colleagues are emphasizing this faceted approach to browsing in pioneering work at the University of Huddersfield (Pollitt et al., 1997).

17. Size of displayed indexes (for print media tools)

This is an issue almost exclusively for printed back-of-the-book indexes. I suggest ways in which students can design such an index retroactively, given a limited number of pages of certain size.

18. Search interface; information visualization; explanation

Great indexing and wonderful indexes are worthless without an effective interface that is easy to use and understand. While we have had centuries of experience and tradition with print displays, we are still beginners in designing effective electronic interfaces. Using principles enunciated by Ben Shneiderman (1998) – give users an overview, then let them zoom in on what interests them – I demand that students provide an overall view of their IR databases on their opening screens, without scrolling in most situations. On this opening screen, they must also introduce the main options for searching: electronic term-matching, alphanumeric browsing, and relational classified browsing of up to three facets simultaneously.

19. Record structure; organization of metadata

Underlying every IR design are records in which all the information about documentary units are described. So students must design a record structure in which every essential piece of information can be readily identified and accessed.

20. Display of full text

Finally, for full-text databases, we often want different types of displays for different purposes. If the original documentary unit was created for print-on-paper media, users should have the option to view it in its original format, but the same text should often be reconfigured to facilitate examination and reading on electronic screens.

Having struggled with these 20 design decisions and all the options available, my students should be ready to begin indexing!

References

Anderson, James D. (1979) Contextual indexing and faceted classification for databases in the humanities. In *Information choices and policies: proceedings of the 42nd annual meeting of the American Society for Information Science*, Vol. 16. Minneapolis, MN and White Plains, NY: Knowledge Industry Publications.

Anderson, James D. (1980) Prototype designs for subject access to the Modern Language Association's bibliographic database. In *Data bases in the humanities and social sciences: proceedings of the IFIP Working Conference* (23–24 August 1979), ed. Joseph Raben and Gregory Marks, pp. 291–5. Amsterdam and New York: North-Holland.

Anderson, James D. (1997) *Guidelines for indexes and related information retrieval devices*. Bethesda, MD: NISO Press (NISO Technical Report No. 2).

Anderson, James D. and Pérez-Carballo, José (2001) The nature of indexing: how humans and machines analyze messages and texts for retrieval. Part I: Research, and the nature of human indexing. Part II: Machine indexing, and the allocation of human versus machine effort. *Information Processing & Management* **37**(2), 231–77.

Bates, Marcia J. (1976) Rigorous systematic bibliography. *RQ* **16**(1), 7–26.

Belkin, Nicholas J., Oddy, R. N. and Brooks, H. M. (1982a) Ask for information retrieval: part I. background and theory. *Journal of Documentation* **38**(2), 61–71.

Belkin, Nicholas J., Oddy, R. N. and Brooks, H. M. (1982b) Ask for information retrieval: part II. results of a design study. *Journal of Documentation* **38**(3), 145–64.

Borgman, Christine (2001) Quote on the book jacket of Ronald E. Rice, Maureen McCreadie and Shan-ju L. Chang, *Accessing and browsing information and communication*. Cambridge, MA: MIT Press.

British Standards Institution (1984) *British standard recommendations for examining documents, determining their subjects and selecting indexing terms*. London: British Standards Institution (BS 6529: 1984).

Cooper, William S. (1978) Indexing documents by gedanken experimentation. *Journal of the American Society for Information Science* **29**(3), 107–19.

Craven, Timothy C. (1986) *String indexing*. Orlando, FL: Academic Press.

Gutwin, Carl, Paynter, Gordon, Witten, Ian, Nevill-Manning, Craig and Frank, Eibe (1999) Improving browsing in digital libraries with keyphrase indexes. *Decision Support Systems* **27**, 81–104.

Hjørland, Birger (1997) *Information seeking and subject representation: an activity-theoretical approach to information science*. Westport, CT: Greenwood Press (New Directions in Information Management No. 34).

International Organization for Standardization (1985) *Documentation – methods for examining documents, determining their subjects, and selecting indexing terms*. Geneva: ISO (ISO 5963-1985 (E)).

Mulvany, Nancy C. (1994) *Indexing books*. Chicago, IL: University of Chicago Press.

Pollitt, A. Steven, Smith, Martin P. and Braekevelt, Patrick A. J. (1997) *View-based searching systems: a new paradigm for information retrieval based on faceted classification and indexing using mutually constraining knowledge-based views*. Huddersfield, UK: Centre for Database Access Research, School of Computing and Mathematics, University of Huddersfield. (`http://www.hud.ac.uk/schools/cedar/bcshci.htm`)

Ranganathan, S. R. (1965) *The Colon classification.* New Brunswick, NJ: Graduate School of Library Service, Rutgers, the State University (Artandi, Susan, ed. Rutgers series on systems for the intellectual organization of information, v. 4).

Shneiderman, Ben (1998) *Designing the user interface: strategies for effective human-computer interaction,* 3rd edn. Reading, MA: Addison Wesley Longman.

Weise, Elizabeth (2000) Web changes direction to people skills. Neatly categorized information requires the human touch. *USA Today*, 24 Jan. p. 1D. (Also available as: Search sites brush up on people skills, at www.usatoday.com/life/cyber/tech/review/crg841.htm)

Wilson, Patrick (1968) *Two kinds of power: an essay on bibliographical control*. Berkeley, CA: University of California Press (University of California Publications: Librarianship No. 5).

James D. Anderson has taught indexing, cataloging, classification and IR database design at Rutgers University (New Jersey) since 1977, where he has also been fighting for equal benefits for lesbian and gay employees almost as long! He is the designer of the CIFT faceted indexing and classification system used by the Modern Language Association of America for its international bibliography and database. Email: jda@scils.rutgers.edu

Retrieval of information can take many forms. Users can express their information need in the form of a text queryâ€”by typing on a keyboard, by selecting a query suggestion, or by voice recognitionâ€”or the query can be in the form of an image, or in some cases the need can even be implicit. Retrieval can involve ranking existing pieces of content, such as documents or short-text answers, or composing new responses incorporating retrieved information. Both the information need and the retrieved results may use the same modality (e.g., retrieving text documents in response to keyword queries), or ...Â In an operational search engine, the retrieval system uses specialized index structures to search potentially billions of documents. Indexing, teaching of See: Information retrieval design. James D Anderson. Places the teaching of (about) indexing into the context of the 20 fundamental attributes of all information retrieval databases or systems, ranging from back-of-the-book indexes to physical and digital libraries, indexing and abstracting services, and online or web-based message/document retrieval services. I do not believe that indexing can be taught. Rules, and the reasons for following or not [Show full abstract] following them, can be presented. Design Information Retrieval. What Info? How to Index?Â We must begin our discussion of design information retrieval by first settling on a notion of information in the context of the design process. Perhaps the most generic view of the design process comes to us from observations of design practice in the field of architecture. Kunz and Rittel [1970] developed the Issue Based Information System (IBIS) as a process model for design based on negotiation, identifying three main componentsÂ 1. Organize cases into short, pointed presentations that teach specific lessons based on particular experiences. 2. Index such stories in terms of design situations they address. 3.