MONTAGUE INSTITUTE REVIEW

ISSN 1554-303X

April 2002

The Economics and ABC's of Indexes

What is the role of the A - Z index in a world of intranets, search engines, and hyperlinks? In this article, we describe the manual indexing process, look at the economics of indexing in print and electronic formats, compare the features of the major types of indexing software, and look at the role of the A - Z index in a corporate taxonomy.

What's an index?

For our purposes here, an index is an alphabetical list of terms that describes the content of a work -- a collection of articles, a book, a report, or an entire web site. Or, to put it another way, an index is "a systematic arrangement of entries designed to enable users to locate information in a document" (see the American Society of Indexers FAQ). A good index helps the reader in three ways:

1. *Identification*. Does a particular topic, issue, person, or organization appear in the document?

2. *Discovery.* Are there related topics that might also be of interest?

3. *Location*. Where do I find information on this topic (e.g. a page reference or URL)?

An index can also serve as the nucleus of a vocabulary -- one of the basic building blocks of a corporate taxonomy.

An old tool in a new medium

Computers have made the task of preparing an A - Z index less tedious. Most professional indexers have replaced cards with computer programs that can store and print an alphabetized list, complete with subtopics, cross references, and page numbers. The raw material, though, is still the printed page -- usually page proofs from a book.

On the other hand, computers -- particularly the Internet and World Wide Web -- have made the indexing process more complex. When the content is 100% electronic or when a work is published in both print and electronic form, new issues arise, such as:

• *Is an index really necessary*? The short answer is yes. Unless you're familiar with both the content and the search engine, you can't be sure what word to enter into the search box. Even if your guess is a reasonably good one, chances are the search engine will give you too many results -- or miss a relevant article. For more on this issue, see "Indexes: An Old Tool for a New Medium" by indexer Kevin Broccoli.

• *How to deal with different formats*? Indexing programs designed for books and other print publications produce page references. But if you're indexing a web site, you want references to individual HTML files. If you're indexing a lengthy electronic document, such as a report or user's manual, you want references to specific headings or paragraphs.

• *What should be indexed?* The focus of the traditional indexing process is a book, which is not only a single physical object but also a cohesive intellectual product. In a magazine publishing environment or on the web, individual articles or web pages may be too short to warrant indexing. Taken together, though, they represent a cohesive intellectual product that should be indexed. The Index on the Montague Institute site, representing 10 years of our published articles, is an example of an indexed collection. Our Briefings and course books, which are published in both print and electronic format, have their own individual indexes.

How to create an index: manual method

The way to create a classic back-of-the-book index involves the following steps:

• *Mark the page proofs*. Read the text and mark the significant words or phrases.

• *Type the terms*. Type the index entries and page references on index cards, one entry per card.

• *Alphabetize the terms*. Arrange the index cards in A - Z order.

• *Edit the terms*. Make headings consistent, add subtopics and cross references.

• *Type and proofread*. Type the completed index as it will appear in the publication.

Each of these steps has numerous sub-tasks (for details, see the Indexing chapter in the Chicago Manual of Style). A key point to remember is that a well-written index includes topics that are implied as well as stated explicitly in the text. It also includes thesaurus features such as subtopics and cross references.

The economics of indexing

A professional indexer charges on the average about $4.50 per page of text (250-300 words per page) to produce a traditional back-of-the-book index. For a typical trade book of 300 pages, that amounts to $1,350. The indexing cost is modest relative to the costs of producing a finished work, which includes hundreds of hours on the author's part plus a substantial investment in editing, proofing, design, and layout by the publisher.

Does the cost of an index justify the value? There's no question that a good index saves time in locating information in a specific work. Multiply the time savings by the reader's hourly rate, the number of times an individual consults the index, and the total number of readers that use the work, and the value quickly becomes a significant number. For long reference works designed to be used repeatedly by multiple people, an index is indispensable.

But that's not the whole story. An index is not just a finding tool; it's an intellectual product in its own right. According to one author who decided to index his own book, the cost of illustrations and indexing was about equal to the royalties he received from the publisher. He felt the investment was worthwhile because the index revealed topics that he had not been aware that he was emphasizing. (See "From the Editor" section in the newsletter of the Australian Society of Indexers.) In other words, an index not only makes research more efficient, it can also make it more effective by revealing new insights.

The economics of intranet indexing

The economics of indexing become more problematic when we move from the concept of a single work (i.e. a book) to the concept of a collection of works (e.g. a corporate intranet). That's because in a collection, the factors in the cost/benefit equation -- the value of the employee's time, the longevity and importance of individual documents, and the number of readers -- are harder to pin down. On the other hand, for individual corporate publishers (i.e. "content owners" such as Human Resources, Marketing, or Legal), the cost/benefit picture becomes a lot clearer. This is one argument for "upstream knowledge management" (a bottom-up approach to content management).

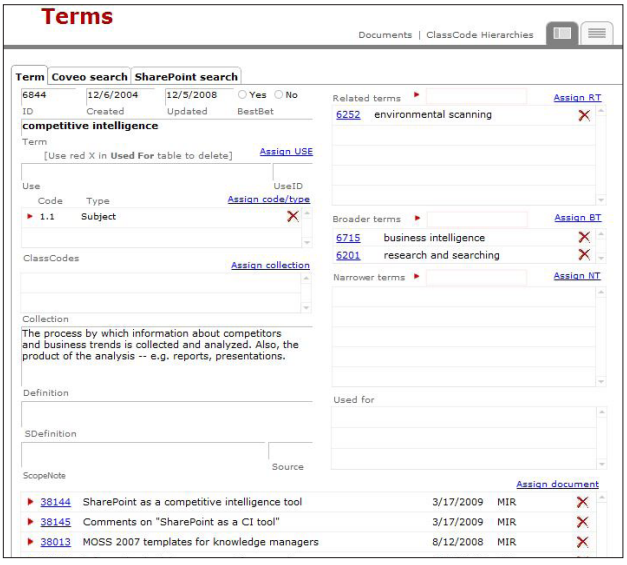
There are two problems with the departmental, or "bottom up" indexing approach:

• how to integrate the indexes of multiple departments;

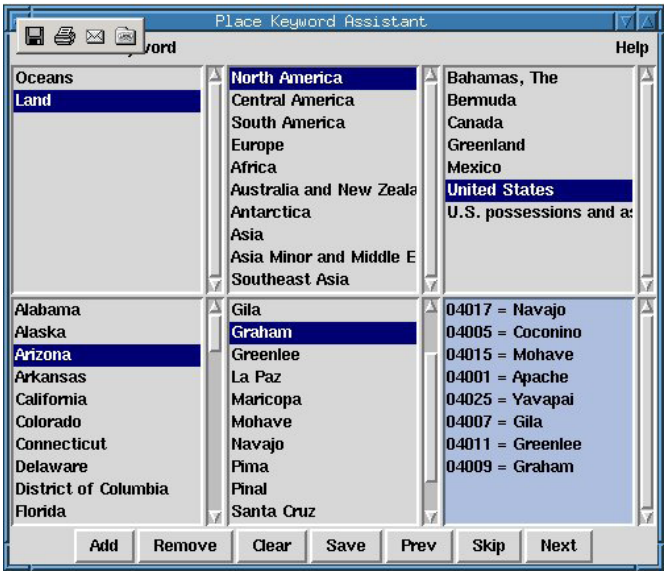
• how to deal with the large number of unindexed documents of unknown quality on corporate intranets.

Fortunately, computers can help in both cases. Indexes prepared for individual publications and departmental collections can be exported to a corporate taxonomy (below, left). In addition, programs can import terms to assist humans in various parts of the indexing process (below, right).can be imported into another program.metadata using keywords from controlled vocabularies"

***Exporting index terms**. The Montague Institute Knowledge Base, like most index management programs, has an export function. The "Export Thesaurus" command creates a series of text files that can be imported into another program.*



Importing index terms. *In this example from the US Geological Service, authors can select terms from an existing vocabulary to help them index their own publications.*

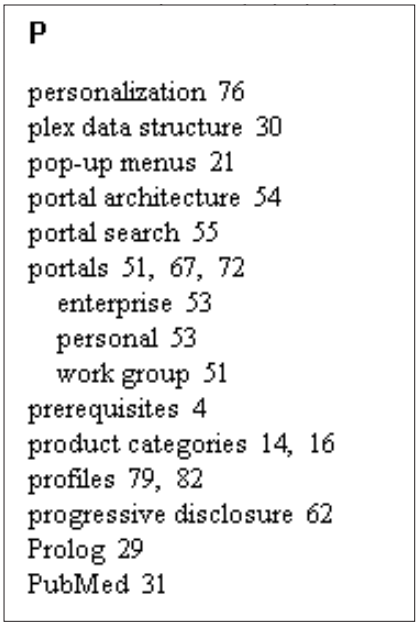


Computer-assisted indexing methods

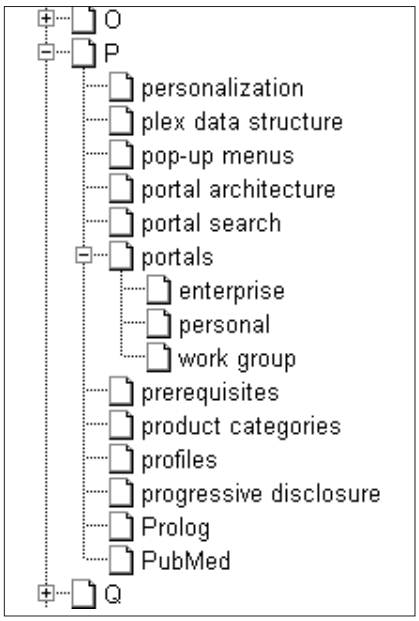
A variety of programs exist to automate various parts of the indexing process. Generally, they fall into the following categories:

1. *Dedicated indexing software.* These programs are used by professional indexers whose primary role is to produce back-of-the-book indexers for commercial publishers. A leading example is CINDEX.

2. *Desktop publishing programs*. Word processing and page layout programs can produce reasonably good back-of-the-book indexes. The process involves manually selecting index terms in the text, inserting special index codes around them, and adding cross references and subtopics. The program will then automatically alphabetize the entries, insert page references, and composes an index section for insertion at the end of the document (see below from Microsoft Word).

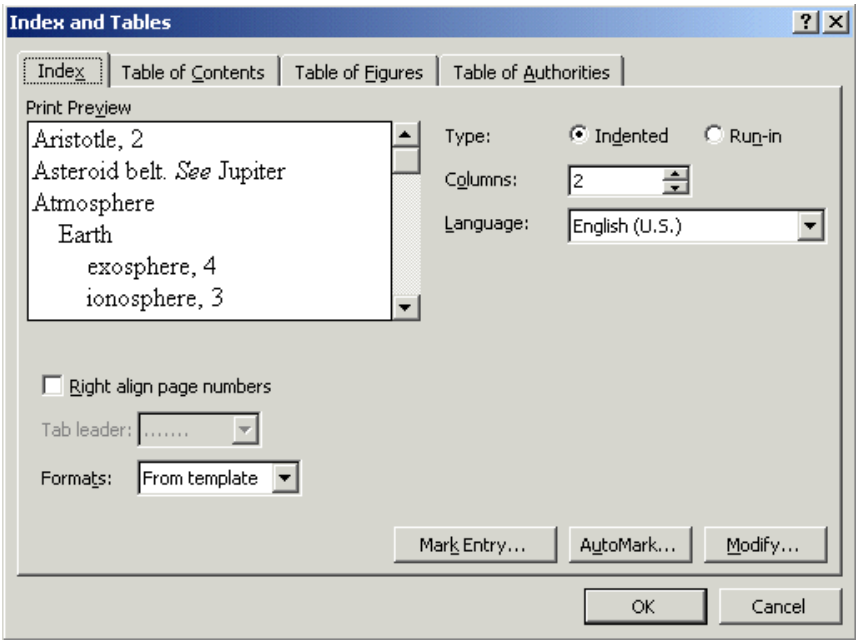


Above: **Before** *A segment of the index as it appears in a desktop publishing program. Microsoft Word produces a similar index. Page references are simple text,not linked.*



After *The index after conversion to PDF using Adobe Acrobat. There is no need for page references, because each term is linked to its occurence in the text.*

Below: The Indexing function in Microsoft Word



3. *Adobe Acrobat*. The full version of the Adobe Acrobat program (not the free reader version) can automatically produce hyperlinked indexes if the original document format supports them. In other words, if we use Acrobat to convert a Word or Pagemaker document to the PDF file format, the page numbers in the index will become "clickable" links in the PDF file.

4. *Relational databases*. Customized database programs (e.g. Oracle, Sybase, SQL Server, Filemaker) are especially useful for producing indexes that cover multiple documents or web pages. If properly designed, database programs can accommodate cross references, hierarchical relationships, and definitions as well as references to the source (a URL or call number). The this method. Because a database reveals the underlying index and thesaurus structure, we also use it as a learning tool in our seminar practicums.

5. *Concept extraction tools*. Programs that identify significant words and phrases in the text can in theory speed up the markup stage of indexing. An example is Copernic Summarizer, a $60 program that extracts concepts and prepares document summaries. You still have to edit the terms, group them into topics and subtopics, create cross references, and link terms to the documents. Some high-end taxonomy programs (e.g. Semio) can do three of these steps -- concept extraction, grouping ("clustering"), and linking (see "Piloting Semio classification software").

The Copernic "Live Summarizer" function identifies index terms and summarizes a document on-the-fly.

The only way to decide whether a computer program can reduce indexing costs is to try it on your content, so we did a simple test. We compared the terms and summaries generated by Copernic with those prepared by our authors. The results are shown in the table on the following page.

Article	Indexer's Terms	Copernic's Terms
OpenURL: a new standard for content integration	content integration hyperlinks metadata OpenURL portals taxonomy standards	database Internet librarians OpenURL references resources server SFX standard web site
Where to find shell companies	company information reverse mergers shell companies	business buslib-l employees joint ventures listed companies phrase reverse merger SDC shell companies website
What is architecture?	architecture authentication databases Inktomi metadata print publishing publishing tools security server side includes style sheet Web publishing	architecture database enlarge folders index integrity membership passwords publishing web site

Copernic versus a human indexer on a sample of three Montague Institute Review articles

None of these programs can do the whole job. Even the most comprehensive (and expensive) require human input and editing. But most of them have both import and export functions, thus making it possible to use programs in combination.

Manual vs. computer-assisted methods

The only way to decide whether a computer program can reduce indexing costs is to try it on your content, so we did a simple test. We compared the terms and summaries generated by Copernic with those prepared by our authors. The results are shown in the table above.

In our experiment, Copernic usually selected terms that were more general than those marked by the authors -- e.g. "business," "Internet," and "resources." General terms like these are useless on the Montague Institute site because they can apply to almost every article, but on a general public site like Yahoo, they might help direct users to the Montague Institute site.

Other test results worth noting include:

• *Number of terms*. The number of author-selected terms varied according to the length and complexity of the article. The number of terms selected by Copernic was set ahead of time in the software options. For the test, we selected 10 terms. With Copernic, there's no easy way to set the number of terms on a per-article basis.

• *Technical terms*. The authors selected terms that were both more technical, perhaps on the theory that users would be more likely to look for them in an A - Z index.

• *Top level terms*. Some terms, such as "print publishing," appear in the document but don't show up in the Copernic list (only the word "publishing" shows up). Yet the authors selected it because it's an important differentiating concept in our content. On the other hand, we've noticed that we tend to overlook legitimate broad categories (see the Subject List on our A - Z Index page), presumably because we can't see the forest for the trees.

Discrepencies. Copernic revealed spelling discrepencies that the authors missed (e.g. "web site" vs. "website").

Summaries. Copernic summaries are not nearly as good as those written by authors. Compare the two summaries on the next page from the "What is architecture?" article.

For us, Copernic is a useful editing tool to suggest additional terms for evaluation by human editors and to catch inconsistencies.

Role of A - Z indexes in a corporate taxonomy

The A - Z indexes prepared for important corporate reference documents are excellent sources of terms for a corporate thesaurus -- a part of the taxonomy structure that is essential in:

• helping users navigate through a list of topics;

• adding intelligence to search engines;

Copernic summary

- The structure of components in a program/system, their interrelationships and the principles and guidelines governing their design and evolution over time (for a definition and related terms, see the Center for Army Lessons Learned thesaurus).
- We draw on ten years of our own experience from 1992 to the present, and describe the evolution of our architecture during three major web site upgrades.
- Good luck if you ever want to find the saved document or revisit it on the web!
- An alternative -- an electronic index using a database program -- involves a little more work at download time but makes retrieving saved files a lot easier.
- The invention of the World Wide Web publishing ultimately led to an architecture based on metadata and the separation of content from format (see below).
- Stage 2: Membership service Adding a membership service meant creating a separate section of our web site for members only and controlling access to it with passwords.
- We wanted a search engine that would combine precision with comprehensiveness by taking advantage of the terms and categories in our Index.
- The Index and Thesaurus database, a part of our Knowledge Base, was loaded onto our web server.
- The end result was a list of documents on our Web site that matched the term, along with a list or related terms.
- Even on a relatively small web site list ours (less than 1000 documents), Google often presented too many hits.
- Three approaches to architecture. There are three basic choices when it comes to selecting an architecture.
- We use two different operating systems, and we try new software vendors frequently as part of our research effort.

Indexer summary

Architecture as it applies to information systems is the structure of components in a program/system, their interrelationships and the principles and guidelines governing their design and evolution over time (for a definition and related terms, see the Center for Army Lessons Learned thesaurus).

As our computing activities become more complex, architecture becomes increasingly important. In this article, we describe the concept of architecture and show how it impacts the cost effectiveness of publishing and information retrieval operations. We draw on ten years of our own experience from 1992 to the present, and describe the evolution of our architecture through four stages:

Making the transition from print to web publishing
Adding a membership service
Separating format and content
Integrating search and browse with a customized search engine
In this discussion, we are concerned with computer architecture in the general sense. The term “information architecture,” which deals with web site design and usability issues, applies to a subset of the topic.

Includes drawings, links to examples of other architectures and related articles.

• integrating the specialized vocabularies of different departments, business units, and geographic regions.

The relationship between the A - Z index of a specific work and the corporate taxonomy is potentially a two-way flow. Terms from the index can be imported into a corporate thesaurus, and thesaurus terms can be used to mark index terms for new publications. Neither of these processes, however, is completely automatic. Human editing is required.

Conclusion

By giving users a familiar, browsable structure of terms and cross references, the A - Z index eliminates a major frustration of full text search engines -- the inability to formulate an effective query. More than just a finding tool, the index is an intellectual product in its own right, capable of shedding new light on a subject.

The value of an index varies with the time value of the user, the number of users, and the frequency of use. The cost is modest compared to the total cost of a professional book. The economics of indexes are harder to calculate on a corporate intranet because the benefits to specific users are harder to pin down. From a quality control and cost/benefit point of view, it’s easier to index departmental collections and then integrate them into a corporate taxonomy.

A variety of software tools are available to reduce indexing costs at each stage in the indexing process. These include dedicated indexing programs used by professional indexers, desktop publishing programs like Microsoft Word, Adobe Acrobat, relational databases, and “concept extraction” tools. Even the most sophisticated and comprehensive tools require human editing.

In addition to helping create indexes, programs can also be used to deploy them for use in other applications (e.g. electronic commerce) or indexing new documents.

The Montague Institute Review *is published by the Montague Institute and edited by Jean Graef.*