Computer-assisted database indexing: the state-of-the-art

Gail M. Hodge

Discusses the state-of-the-art in computer indexing, defines indexing and computer assistance, describes the reasons for renewed interest, identifies the types of computer support in use using selected operational systems, describes the integration of various computer supports in one database production system, and speculates on the future.

Computer-assisted database indexing, particularly automatic indexing, has been discussed for many years in the information science and database production arenas. It has had proponents in both commercial and academic circles. Automatic indexing has met with staunch opposition as intrusion on an intellectual activity, reserved for humans, and has been equally touted as a means of raising the human indexing effort to new levels. Old debates aside, there is renewed interest in computer-assisted indexing. This interest is not merely theoretical; there are many operational systems in place supporting the work of major database producers.

Definitions of indexing and automation

According to Frederick Lancaster[1], 'The main purpose of indexing . . . is to construct representations of published items in a form suitable for inclusion in some type of database. The database of representations could be in printed form, in machine-readable form or in card form'.

Indexing in the context of this paper is more limited, referring to the representation of the subject content of a source document which results in access points. Access points that result from subject indexing are distinguished from those that result from descriptive cataloging, such as authors and journal names. Subject indexing provides access points to the document that may not appear explicitly in the document itself. These additional access points may be the result of the addition of terms not found in the record, or may result from the standardized presentation of the natural language appearing in the document.

This definition could stand equally for back-of-the-book indexing or database indexing. However, database indexing generally provides access to a large amount of literature—numerous individual books or journal articles—by imposing a subject-base structure, or indexing scheme, on the content of the resulting database. The basic structure of this indexing scheme is intended to apply to the general subject matter of the database contents over a long period of time.

Automated support to indexing could be defined as a robotic arm used to turn the pages of an indexer's printed thesaurus. While that may be an important productivity aid in the right setting, ergonomic issues such as robotic arms or computer-controlled environments are not the focus of this paper. Rather computer support refers to the use of some type of computer to aid, directly or indirectly, the intellectual and clerical tasks of subject analysis and access point assignment.

Investigations of machine-aided indexing, automated indexing, and computer-assisted indexing date back to the late 1950s. However, many of the ideas proposed for text analysis and automated indexing proved too simplistic for the complexity and richness of natural language. With the advent of large online retrieval systems the research emphasis shifted to improvements in information retrieval—methods for extracting from the database, rather than to indexing—methods for tagging the data going into the database. Even in a recent *Monitor* editorial, Harry Collier laments, 'little is being done in this subject [automatic indexing] and, when things are done, the scale is small'.[2]

Collier may be accurate in his assessment, but my interviews with a few database producers indicate that this is changing, at the impetus of the producers themselves.

There are several reasons for renewed interest in automated indexing on the part of database producers. Chief is the less than favorable economics of database production.

The economics of database production

In July 1991, the National Federation of Abstracting and Information Services, an association of database producers, reported that 'belt tightening has become a way of life, the market is producing diminishing returns, and dire-consequence economic headlines abound.'[3] In the same issue, NFAIS reported

Paper presented at the American Society of Indexers' conference, 'Indexing, providing access to information: looking back, looking ahead', Alexandria, Virginia, 21 May 1993

on the results of a survey on cost containment among secondary services. Of the NFAIS member organizations that responded to the survey, 84% were taking additional steps to deal with the current economic situation. The producers are trying to provide the same or more services to their users with fewer resources.

At the same time, database users, particularly libraries supporting academic and research institutions, are caught in a similar bind. In the same July 1991 *NFAIS Newsletter*, Fayen reports on the problems of library and information center management in a time of '. . . fewer resources, fewer staff, and even (perhaps) fewer patrons demanding fewer services'.[4]

If information is considered a luxury that can be sacrificed, or if budgets for information collection and dissemination are curtailed due to the corporate economic climate, the services will be used less often, resulting in a vicious cycle for all parties in the information chain.

In response to the current economic situation, database producers are evaluating their internal operations—what they do and how they do it. Indexing is an integral part of most database production systems. It is often the producer's bread and butter, and along with providing access to a large composite of material, the indexing overlaid on that material is the added value provided by the producer.

However, indexing historically has been a very labor-intensive function, perhaps the most labor-intensive one within a secondary information service. According to the NFAIS cost containment survey, 56% of database producers surveyed were reducing costs by automating their systems to greater or lesser degrees. Of those who sought to reduce costs through automation, six cited cost savings through automated indexing.[3]

Add to the current economic situation concern for indexing quality, problems of indexer retention, and the need to renovate aging production systems, and it is easy to see the reason for renewed interest.

Computer-assisted support continuum

The renewed interest on the part of database producers was brought to the fore by a meeting at the US Department of Energy's Office for Scientific and Technical Information in September 1990. Dr Jessica Milstead of the JELEM CO. moderated a meeting of 13 sci-tech database producers.[5,6] This meeting stimulated my writing a book on the use of computer assisted indexing in the production systems of 23 database producers in the government, commercial and not-for-profit sectors.[7]

To analyze the operational systems, I developed a computer-assisted indexing continuum. This runs the gamut from no automated support at all to full automatic indexing, with a large area in between for various types of computer-assisted indexing.

No Computer Support	Computer Support to Clerical Activities	Computer Support to Quality Control	Computer Support to Intellectual Activities	Automatic Indexing

Figure 1

The left end of the spectrum is no computer support to indexing; on the far right is automatic indexing with no human review. In between are computer support to purely clerical activities such as index term input and index term location; computer support to the quality control of index terms; and support to intellectual activities, including index term selection, weight assignment and index term order.

Computer support to clerical activities

Clerical activities include locating the index term, generally in a thesaurus or other vocabulary aid, and the entry of the term into machine-readable form for the document being indexed.

Online indexing aids that help the indexer locate the term to be used may be of several types—thesauri, controlled vocabulary lists, scope notes, and support files. Many database producers provide vocabulary aids connected to online thesauri. The thesaurus acts as an online reference tool that can be queried faster than a printed resource. For example, the FIBRE system at the International Nuclear Information System in Vienna, Austria allows the indexer to select a natural language term from the title or abstract, and the system searches the permuted list of words from valid thesaurus terms for a match. The vocabulary often is supplemented by scope notes and instructions on the use of the index term. In the new PC-based indexing system at the American Institute of Aeronautics and Astronautics (AIAA) the thesaurus 'sections'—valid terms, related terms, and scope notes—are available on separate screens with functions keys between them.

Numerous methods exist for eliminating/reducing the clerical (data entry) aspects of the indexer's task. Many organizations with thesauri or controlled vocabularies allow the indexer to 'cut-and-paste' terms from the online thesaurus into the document without rekeying. The system at the AIAA automatically inserts the term into the indexing record when it is selected from the thesaurus. Analysts at Chemical Abstracts Service (CAS) select a term and look it up in the authority file in one window in an X-Windows environment. The controlled term can be 'copied and pasted' from the authority file into the document in the main window.

Computer support to quality control

Validating the indexing is done in both batch and online (interactive) modes. Batch processing still is prevalent; it is used alone and in combination (as a doublecheck) on the interactive validation. Both BIO-

SIS and Chemical Abstracts Service (CAS) use overnight batch processing to perform quality control checks on the terms input by the indexers. Reliance Medical Information (RMI) uses a variation on batch validation. The subject headings are input by health professionals and checked interactively; when the diskettes are returned to the RMI staff, the subject headings are rechecked against the subject headings on the CD-ROM drive.

Most online systems provide some level of interactive validation, the simplest being character set and code checks. Systems that allow the indexer to key index terms or surrogate codes generally provide validation of the terms or codes. The system in use at Petroleum Abstracts, University of Tulsa, returns candidate preferred term(s) when an invalid term is input.

Systems that use natural-language index terms may use spell-check dictionaries. These may be created from previously indexed documents based on a frequency threshold, or be bought by purchasing specialized subject dictionaries from external sources. BIOSIS uses a spell-check program and dictionary written in-house to validate the natural language terms in its BioBusiness database.

Computer support to intellectual activities

The next milestone on the continuum is computer support to intellectual decisions, i.e., the selection of terms. Computer support to intellectual decisions in itself contains a broad range of support tools—from prompts to remind the indexer that certain indexing content is required to computer analysis of text data elements to provide candidate indexing.

Aids that prompt the indexer for particular types of index terms may be presented on a standard screen (template) or interactively as the indexing of a document progresses. One of the most notable examples of a prompting aid is the National Library of Medicine (NLM)'s check tag system. General index terms such as bibliographic format and age/gender of an organism are presented on the screen as prompts (check tags). Some check tags are required based on the subject of the article or previously entered index terms. For example, if a clinical report includes the term 'patient,' the check tag for age/gender is required.

Indexing assigned by text analysis is accomplished most often in a batch mode, because the analysis is too time-consuming or complex to handle interactively in real-time mode. The American Petroleum Institute analyzes the titles and abstracts of documents to assign thesaurus terms using a rule-based expert system. The International Nuclear Information System and FIZ Karlsruhe, a German producer of physics and chemical databases, use expert systems based on the statistical correlation between significant natural language terms and assigned index terms in previously processed documents. Information Access Corp. (IAC) evaluates the titles and significant cues, such as SIC numbers and the indexing received from newswire services, to map to IAC indexing terms.

Candidate terms are less often created interactively as the indexers work on each document. However, examples do exist. IAC's indexing support is based on authority files stored on CD-ROM. The authority files provide related data, such as SIC codes, interactively when particular index terms are added. The system in operation at NASA STI Program Center for Aerospace Information provides candidate controlled terms from the NASA Thesaurus in both batch and interactive modes. Semantic units are parsed from titles and abstracts; these units are matched against the knowledge base of 11,000 natural language terms connected to thesaurus terms.

Access to previously indexed documents

Access to previously indexed documents can be viewed at several points along the continuum, but is aligned most closely with aid to the intellectual process. Previously indexed documents can provide consistency across time by presenting real document-specific examples of the application of indexing terms and policies. Previous documents may also be valuable when the indexer is trying to understand one that contains unknown terminology, or unexplained acronyms or abbreviations.

Many indexing organizations provide online access to previously indexed documents. Producers of electronic databases may provide access to previous documents through in-house text retrieval systems or through external online search systems. The newest medium for this type of automated support is CD-ROM. CD-ROMs originally produced for customers also can be used by the indexers.

Automatic indexing

Several organizations, including BIOSIS, the Defense Technology Information Center (DTIC), FIZ Karlsruhe, the Russian International Centre for Scientific and Technical Information (ICSTI), INSPEC/Institution of Electrical Engineers, and Institute for Scientific Information (ISI) use automatic indexing of documents without human review. The number of organizations is greater than expected, considering the effort required to build systems that are capable of this level of performance. However, in all cases the automatic indexing is done on only a part of the database—either a specific subject, document type, or access point.

Automatic indexing may be implemented for a specific subject or document type. For example, until 1992, ICSTI analyzed the Russian language titles and abstracts of the technical report and dissertation documents to assign indexing terms. DTIC uses automatic indexing to assign thesaurus terms to only the management related literature in its database.

Automatic indexing may provide additional, derived access points. FIZ Karlsruhe's automatic chemical name-derivation system locates chemical formulas, places them in a special field, and inserts the appropriate base elements as a searchable prefix to the formula. This program is used for other databases mounted on STN International. The American Petroleum Institute's computer expansion for Markush-type representation allows the indexer to enter the sub-elements in a series of lists, and the computer explodes these lists into the appropriate combinations. The system reduces the time spent indexing some patents from hours to minutes.

Automatic indexing may be the only cost-effective way to re-index a database's backfile. INSPEC used automatic indexing to change the indexing applied to old records that could not possibly be brought up to new standards manually. However, INSPEC does not use automatic indexing for routine production.

Certain indexing schemes more easily accommodate automatic manipulation. ISI's co-citation indexing used in Science and Social Science Citation Indexes lends itself to statistical and bibliometric analysis. 'Research Fronts' annually regenerates clusters of core papers and related papers based on a variety of statistical algorithms. 'Keywords Plus' automatically indexes documents with additional keywords from the titles of cited references. 'Related Record' connects records based on shared citations. These product enhancements are based on deeper automatic analysis of the relationships between documents, and their keyword indexing, already established through co-citation indexing.

For purposes of discussion I have defined an automatic indexing continuum, which stretches from support to clerical activities to automatic indexing without human intervention. However, this continuum is a somewhat arbitrary construct. Database producers may appear in one or more places on the continuum at the same time, depending on each organization's needs, resources, indexing environment, the material being indexed, and production system constraints.

One company on the continuum

The integration of computer-assisted supports is best illustrated by looking at the indexing system in a single company. I will use BIOSIS as an example of how various computer-assisted supports to indexing are used to improve quality, reduce processing time, and generally support the effort of indexing a large database. (BIOSIS is the world's largest database producer in the areas of life science and biomedicine. Approximately 600,000 documents from books, meetings, journals, and patents are indexed annually.)

The indexing scheme used to produce BIOSIS Previews and its equivalent print products (*Biological Abstracts* and *Biological Abstracts/RRM*) is based on a limited, high-level, controlled vocabulary and natural language keywords. The indexing scheme has four parts: a controlled thesaurus of general biological concepts (concept codes); a companion controlled thesaurus of general taxonomic or organism-related terms (biosystematic codes); controlled keywords; and natural language terms.

Computer-assisted indexing is used in a number of areas—clerical support, support to the intellectual process, support to quality control, and automatic indexing.

In 1991, BIOSIS installed a PC-based indexing system called 'SYstem for iNDEXing' (SYNDEX). SYNDEX is written in WordPerfect 5.1 macros and supplementary C programs. The design team, made up primarily of indexers, was sensitive about overengineering the design; it tried to give indexers the freedom to use the full power of WordPerfect (without windows). SYNDEX currently runs on stand-alone PCs running DOS. Records are uploaded onto the mainframe from floppy diskettes.

SYNDEX's major emphasis has been on speed and consistency by providing support to clerical and intellectual activities. The added keywords, concept codes, and biosystematic codes can be input with aid from several authority files. The indexer can select the code/term from an authority file without rekeying, or use the authority file to validate the entered data. If the tie between keyword and code is strong, the system supports the intellectual effort by supplying the code automatically when the term is selected. One of the most important files lists genus epithets (abbreviations). If an epithet occurs in the paper, and the indexer does not know the appropriate full genus name, the indexer enters the epithet, and the system provides the full genus name from the Genus Epithet File. The expansion can be added to the index record by a single keystroke.

In addition to the aids provided through the PC-based SYNDEX software, the indexers have online access to previously indexed documents and online authority files on the IBM mainframe. Up to 18 months of the BIOSIS Previews database are available under the IBM-STAIRS search software. Indexers search the in-house BIOSIS Previews database to solve the intellectual problems of indexing. For example, an indexer might access several papers by a particular author to clarify past use of a seemingly obscure abbreviation. Several online authority files also are available. An online drug file allows the indexer to obtain information on the generic, chemical, and trade names of pharmaceuticals encountered in the literature, as well as the mechanisms of action of these compounds. A similar file exists for other chemical terms.

Access to previously indexed documents and to authority files is limited, however, because the files are available on the mainframe, not on the PCs on which the indexers actually work. These files will be inte-

grated with the PC indexing system in the planned new production system.

The indexer can reuse intellectual effort by reusing the indexing. Since many papers from the same meeting carry identical indexing terms, the indexer can replicate terms from a record into subsequent records with a single function key. In addition, the indexer can create a master record, carrying the indexing that is universal to a set of meeting papers, and then duplicate (clone) that record for each subsequent paper to save keystrokes and ensure consistency.

The SYNDEX System provides additional computer supports for quality control. The WordPerfect spell checker has been enhanced with several commercial dictionaries and with highly frequent BIOSIS terms to spell-check the natural-language terms.

After the indexed records have been selected for BIOSIS Previews, automatic indexing is used to add Chemical Abstracts Service (CAS) Registry Numbers (RNs) to a version of Previews (BP/RN) available on STN International. CAS RNs are valuable access points for BIOSIS users interested in chemicals, drugs, enzymes, or biosequences. The addition of RNs to the entire BIOSIS retrospective file, back to 1969, was accomplished through automatic indexing.

The CAS RNs in BP/RN are assigned by an expert system that analyzes the natural language text using a series of authority files to identify chemical names. Potential chemical names are identified using a series of natural-language processing techniques including flagging of terms that may surround chemical names, special handling of punctuation variants, elimination from the matching process of a vast stopword file (those that have previously been unmatched), and a series of post-processing rules to assign derivatives and questionable assignments [multiple RNs (chemical structures) with the same name]. High-frequency chemical names are assigned RNs from a file maintained by BIOSIS. Less frequent RNs, and RNs for new chemical names, are assigned by sending 'unknown' words and phrases to Chemical Abstracts Registry Services, where the words are matched against CAS' Name File. The BIOSIS and CAS-supplied RNs are merged with the Previews records and loaded on the STN International system.[8,9]

The maintenance of authority files is important in the BIOSIS RN system, since the RN assignments on the individual papers are not reviewed. The human effort is devoted exclusively to the maintenance of the authority files and rules. This effort requires approximately 5 hours per month, far less than the time that would be required to evaluate and manually assign Registry Numbers to 540,000 papers.

The Registry Numbers assigned to BIOSIS Previews/RN are assigned without human review. This has proven extremely successful for this limited application, where intense analysis of the text is not needed.

Future trends

Using the computer to support the clerical, intellectual and quality control aspects of indexing is blended in an operational indexing environment at BIOSIS and in the operations of many other database producers. The future holds the promise of ever increasing efforts in the areas of computer-assisted support tools, including more use of automatic indexing with little or no human review. Based on the pragmatic needs of database producers, including new information entrepreneurs, academic and commercial researchers and software developers will produce increasingly flexible and efficient systems. The future will see the continuous addition of computer-based tools in the indexer's toolkit, tools that will benefit the indexing activities of all of us.

References

1. Lancaster, F. W. *Indexing and abstracting in theory and practice.* Champaign, IL: University of Illinois Graduate School of Library and Information Science, 1991.
2. Collier, H. EUSIDIC in Seville. *Monitor* **129**, 1991, 103.
3. Strategic thinking in a slumping economy. *NFAIS Newsletter* **33** (7) 1991, 92–4.
4. Fayen, E. G. Managing in tough economic times. *NFAIS Newsletter* **33** (7) 1991, 89–90.
5. Milstead, J. L. *Methodologies for subject analysis in bibliographic databases: background paper and report of meeting sponsored by International Atomic Energy Agency and Energy Technology Data Exchange*, September 1990.
6. Milstead, J. L. Methodologies for subject analysis in bibliographic databases. *Information Processsing & Management* **28** (1) 1992, 407–31.
7. Hodge, G. M. *Automated support to indexing.* NFAIS Report Series #3. Philadelphia, PA: National Federation of Abstracting and Information Services, 1992.
8. Hodge, G. M, Nelson, T. W. and Vleduts-Stokolov, N. *Automatic recognition of chemical names in natural language texts.* Paper presented at the 198th American Chemical Society National Meeting, April 1989.
9. Hodge, G. M. *Enhanced chemical name identification algorithm.* Poster presented at the SciMix Poster Session: Fourth World Chemical Congress, August 1991.

Gail Hodge is former Systems Administrator, BIOSIS, Philadelphia; now Director of Operations & Analysis, US National Aeronautics and Space Administration Center for AeroSpace Information, Maryland.

Non-indexer in fiction

Having completed her biography of Edith Wharton, Loretta, heroine of Joan Smith's *Don't leave me this way* (Faber, 1990), proudly tells a friend enquiring how many words she has written. 'Ninety-eight thousand, not including notes and the bibliography, of course. And there's still the index, but I think I'm going to pay someone to do that'.

—That's the sort of author we *like* to hear about.