Title

AN INTERACTIVE INDEXING-EDITING SYSTEM FOR DOE TECHNICAL INFORMATION CENTER

Authors
Cerny, B.A.
Lawrence, J.D.

Publication Date
1982-02-01

LBL-14049

Lawrence Berkeley Laboratory

UNIVERSITY OF CALIFORNIA

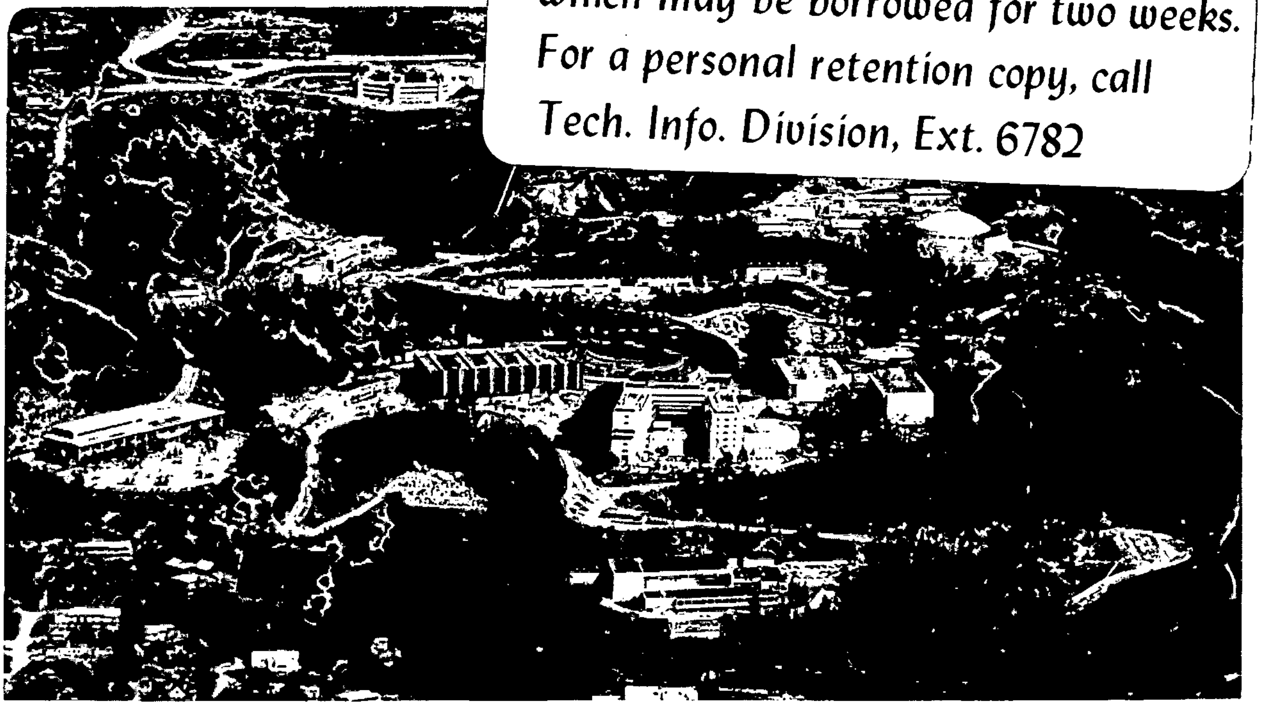
Engineering & Technical Services Division

Submitted to the American Society for Information Science Mid-Year Meeting, Knoxville, TN, June 13-16, 1982

AN INTERACTIVE INDEXING-EDITING SYSTEM FOR DOE TECHNICAL INFORMATION CENTER

Barbara A. Cerny and J. Dennis Lawrence

February 1982



Prepared for the U.S. Department of Energy under Contract W-7405-ENG-48

DISCLAIMER

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor the Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or the Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or the Regents of the University of California.

LBL-14049

AN INTERACTIVE INDEXING-EDITING SYSTEM FOR DOE TECHNICAL INFORMATION CENTER

Barbara A. Cerny and J. Dennis Lawrence

Lawrence Berkeley Laboratory
University of California
Berkeley, CA 94720

This work was supported by the Director, Office of Energy Research, Office of Basic Energy Sciences, Materials Sciences Division of the U. S. Department of Energy under Contract No. W-7405-ENG-48

An Interactive Indexing-Editing System for DOE Technical Information Center

Barbara A. Cerny and J. Dennis Lawrence

Lawrence Berkeley Laboratory
University of California
Berkeley, CA 94720

Abstract

A system to provide computer assisted aids to document indexing for the Department of Energy Technical Information Center Energy Information Data Base is being developed. The goal is to allow more efficient processing of documents and document surrogates by using interactive computer techniques rather than the current combination of manual indexing and batch computer processing. It draws on automatic indexing techniques, but includes the indexer as an integral element of the system; he is the decision-maker while the computer provides "clerical" support.

This approach is transferable to other large production systems.

1. Introduction

The DOE Technical Information Center (TIC), Oak Ridge, Tenn. serves as a collection, processing, and distribution point for all energy-related information. The information comes in from both national and international sources on magnetic tape and in printed form. Items judged to be of interest to DOE and its contractors are processed and become part of the DOE Energy Information Data Base (EDB). [1]. From this data base, TIC produces abstract journals, bibliographies, magnetic tapes for DOE/RECON, an interactive retrieval system, [2] as well as tapes for the International Nuclear Information System (INIS), SDI and other technical services.

The point of contact with TIC's processing discussed in this paper, is in the area of document indexing for EDB preparatory to entry into the RECON system. Currently this data base includes more than 1,000,000 citations and more than 15,000 items a month are added to it. The documents, or document surrogates are processed with a combination of manual indexing and batch computer processing. Approximately 65% of the citations come in on magnetic tape from American Institute of Physics, Engineering Index, INIS and other sources, already indexed to some degree, while the remaining 35% are printed documents.

The problem posed with regard to this document processing, was how to use state-of-the-art interactive computer techniques to aid the indexers in performing their jobs more effectively, hence reducing the backlog of work caused by the large volume of documents passing through the system. Initially, there was to be as little disruption of production as possible; the new system would work in parallel with the old until it was refined enough to replace the current methods. Hence it was decided to work at first with only the 65% of the document surrogates that arrived on tape. This paper is a report on the system to date; it is a work in progress which is drawing from current research spanning information science to linguistics, to provide a man-machine interface for document processing applicable to large production systems. It is not an automatic indexing system intended to replace the indexer; rather, the indexer is considered to be an integral element in the system.

From a system design perspective, there are three primary considerations:

A. The computer programs are to be written in a portable language, within certain constraints. First, there is a great deal of string or text manipulation and almost no "number crunching". Second, the system has to run on a PDP10 computer. Third, there are authority lists, such as the thesaurus, already in place and used in the production work at TIC. The access routines, written in PDP10 assembler, must be integrated into the new code. These machine dependent files place a limitation on the portability of the system, regardless of the language it is written in.

The language of choice is RATFOR ("Rational Fortran" as documented in Software Tools [3]), a preprocessor that translates into Fortran. Since Fortran is such a ubiquitious language, and RATFOR itself is enjoying wider distribution, the work necessary to port this system should be as mimimal as possible.

B. User friendliness is vital, yet in a production environment the demands on the machine have to be kept to a minimum. If one looks at interactive programming style and theory, the most demanding from all perspectives is natural language dialogue, the goal of much Artificial Intelligence research, while at the other end of the spectrum is the highly formalized, usually terse prompts and responses that suffice as man-machine communication in a structured business environment. TIC's system lies somewhere between. It is menu-driven so that it minimizes typing demands on the part of the indexers while it avoids the heavy computer demands of a natural language parser. Yet it allows the indexers great flexibility in moving through the universe of indexing concepts and steps.

C. The clerical demands on the indexer are to be reduced and his decision making enhanced. Whether an indexer has access to terminal display of documents as opposed to hard copy, a keyboard instead of a data-entry sheet and a pencil, it is only the clerical function, the thumbing through printed authority lists for suggestions and verification, the routing and rerouting of documents by hand that are replaced in this system.

2. The Indexing Process

There are two levels at which the design of such a system could be approached. First, the indexing process has to be defined at the surface level - what steps and rules do the indexers follow in processing a document or document surrogate. The second question has to do with cognitive processing, with the intellectual devices by which an indexer "understands" text and translates it into the metalanguage of a controlled vocabulary. Since the primary goal is to construct a working on-line interactive system to speed production, it seemed that a rigorous definition of the process at the surface level would allow the definition of modes of operation or dimensions in which the indexers work. The cognitive processes invoked in each mode, the mechanisms by which "intelligence" could be built into the system, could then be addressed at a later date using these basic, functioning programs as building blocks of this inquiry.

Briefly, then, following and expanding upon Digger [4] the intellectual effort in TIC's indexing procedure involved is as follows (not necessarily in order):

A. Scanning of the text (or abstract), title, author, and other cataloging material.

B. Assessing the nature of the document.

C. Identifying the concepts.

D. Relating the concepts to user requirements.

E. Generating an abstract, if necessary.

F. Generating a title augmentation, if necessary, to replace the title with one that has information content, or to allow retrieval when a descriptor pair and title are combined into a subject entry [5].

G. Selecting the concepts to be indexed.

1) Forming concepts with respect to subject categories, and assigning one of 861 subject categories (average (1.3)) [6].

2) Translating concepts into thesaurus descriptors [7]. This includes descriptor splitting to prevent false coordination for users and descriptor flagging to identify descriptors as main headings or qualifiers [5].

H. Checking the work.

The modes of indexer operation that were defined arose from this breakdown, coupled with the goals of user friendliness and the minimization of clerical and verification tasks. This categorization is for design conceptualization only - the indexer is aware only of one menu-driven procedure.

A. Display mode. This consists of:

1) A menu display, giving a choice of operations to be performed on elements within records, or actions to be performed on the records the themselves.

```
Enter a field/operation
t - - - - - (primary title (A))
s - - - - - (subtitle (A))
f - - - - - (file selected for)
c - - - - - (categories)
k - - - - - (keywords)
a - - - - - (abstract)

* - - - - - (display entire record)
+ - - - - - (add a new field)
- - - - - - (delete an existing field)
# - - - - - (edit a field other than the above)
@ - - - - - (reroute the record)
^ - - - - - (reject the record for lack of EDB code)
% - - - - - (delete the record)
? - - - - - (help)
$ - - - - - (terminate the run)
```

2) A record element or field display. The menu provides single keyboard stroke access to the primary fields that are most frequently operated on such as abstract, keyword, category, and title fields. There is an option to display any field in the record, or the entire record.

B. Edit mode. Any element or field in the record can be edited by a powerful screen editor. This editor allows cursor control and text manipulation similar to that of a word processor. This is the heart of the indexing procedure. Keywords can be entered or deleted from the record, spelling corrected, words capitalized, etc.

C. Verification mode. This is automatic. If an indexer enters a keyword or subject category it is verified against an on-line authority list. If it is not an allowed entry a message is displayed to that effect.

D. Help mode. At any point, one of two forms of help can be requested - help with the editor commands or help with main program control.

E. Link mode. When the title and abstract are displayed through the display mode, an automatic check is made of the stem of each word against the thesaurus. If a match is found, the word is highlighted on the CRT screen. It is then possible for the indexer to trace associations to that word with his link map from the permuted index of the thesaurus or from the thesaurus itself. If he enters the thesaurus display, terms associated with a term in the title or abstract such as broader, narrower or related terms, definitions, etc., are offered to the indexer for his perusal, (Figure 1) [8]. It is possible to move through this concept space by keyboarding the next term to be examined and, eventually, a touch of a term with a light-pen will enter this word as the next display.

Since an abstract is a condensation of a document, the vocabulary in the title and abstract can only indicate the choice of words an author used; it will not necessarily reveal all or even most of the concepts and ideas in the document. It is the indexer's problem to find his way from the natural language of the title and abstract to the allowed thesaurus terms, and the highlighting will give an entry into this process. It was found in a similar experiment mapping text into the INIS controlled vocabulary [9] that after morphological analysis only 10% of the title and abstract text matched the thesaurus. Another experiment [10] showed that only 40% of the assigned descriptors contained in the text title and abstract of Petroleum Abstracts matched the Exploration and Production Thesaurus that was used to index that publication. Matching in neither case provided a substitute for manual indexing, but with the inclusion of the indexer as a component of the system, matching and highlighting can provide a significant entry point to the indexing process.

F. Statistics mode. Program modules are being written to collect and display statistics about the movement of records through the system (described below) as well as about the movement through the different modes. The first case is a management or supervisors tool to keep track of the "time and motion" of the indexing process. The second case will be used as feedback to the designers and programmers to refine the system.

3. File Management

Since TIC's current manual and batch procedures for dealing with document surrogates that come in on magnetic tape are being transferred to computer control, a description of these procedures will provide a baseline for examining the new system.

A. As the tapes come in, they are run through a batch process that checks elements against authority lists, notes errors, formats the records for easy manual scanning, and prints these reformatted records.

B. The printout is routed from indexer to indexer, each of whom works with the records pertaining to his speciality. Indexing is done by subject, with each indexer, or subject specialist being responsible for certain areas as defined by the EDB Subject Category descriptions [6].

C. After corrections have been made and categories, index terms and abstracts added as needed to all the records that are relevant to energy, the printout goes to a descriptive cataloguer who enters these changes on-line into an image of the record stored in the PDP10.

D. After checking and reindexing, if necessary, a magnetic tape is generated for entry into the RECON system.

From a file management perspective, there are three phases from the entry of an input tape into the system to the generation of a RECON tape, (see Figure 2). Since this system is designed on a PDP10, we shall speak of "areas" or "directories" in which files consisting of document records are stored. There must be one area for each indexer, one for each Supervisor, plus a Master Area. The guiding philosophy is that a record, except for archival backup, exists in only one file at a time but it can be routed easily to another file depending on the action to be taken on it.

This procedure visualizes several different "roles" for people. A person may fit different roles at different times; the description by role is merely an attempt to keep task categories separate. These are:

A. Indexer. The person who operates on the records correcting errors, completing the index terms, and creating an abstract if necessary.

B. Supervisor. The person who supervises the indexers and who may wish to examine their work.

C. Master Operator. The person who manages the files in the Master Area and runs the programs in Phases I and III.

Referring to Figure 2, the phases, files they generate, and areas of operation are as follows.

A. Phase I. The SORT program This is a batch routing of records from the original tapes into the subject files in the Master Area. The 861 subject categories in EDB are divided hierarchically into 3 levels. There are 40 first level categories such as Coal, Physics, Fossil fuels, etc. and indexing responsibility divides at this level. Hence the routing of documents is to one of these 40 categories. This is done in one of two ways.

1) On some tapes, EDB categories have previously been assigned, so it is a trivial matter to route the documents. In other cases, for example, INIS tapes, there is a mapping between some INIS and some EDB categories, so many of the documents can be routed automatically. Those whose categories do not map onto EDB must be routed by an indexer from an "unknown" file.

2) If there is no such information, then a part of the preprocessing will be a statistical analysis of the words in the title and abstract as compared to a training set composed of the past history of vocabulary in the RECON database.*

*This research, in the realm of automatic classification, will be described in a later paper. Briefly, the past history of vocabulary usage in each of the 40 categories is being built into an inverted index. We are experimenting with algorithms for predicting category based on this file and the text of the abstract in question. A preliminary experiment with 18 first and second level categories from the first level categories "Coal" and "Physics" and a small training set of 100-300 documents per category gave encouraging results: 80% of the documents analyzed were correctly categorized.

B. Phase II. The TICEDT program This is the heart of the system. It consists of an interactive program coupled with a screen editor that allows the indexer or supervisor to access the records in the subject category files and operate on them.

1) Indexer. In terms of file manipulation, when an indexer logs into his area, one of two procedures occur.

a. If he had terminated a session in the middle of a file, he will automatically be returned to where he left off.

b. Otherwise, he is prompted for the category he wishes to work on. His response causes a file to be moved from the Master Area to his area, and archived.

As he works on records, he has the options of

a. Deleting a record completely from the system.

b. Rerouting to another category file if he feels the record would be better handled by another subject specialist.

c. Editing and Indexing the record and placing it in a SAVE file to be collected later for incorporation into RECON.

2) Supervisor. When a supervisor logs into his area, he is asked for the indexer whose work he wishes to check. This causes the SAVE file from the indexer's area to be moved into the Supervisor's area. The Supervisor can then proceed to edit this file.

C. Phase III. The APPEND program

This is a batch collection procedure whereby all the indexed records are stored in master files to be made into a RECON tape. At the end of a day, or specified period of time, all the completed records, the SAVE files, are collected from each Indexer's and Supervisor's areas and stored in the Master Area and are deleted from the Indexer and Supervisor Areas.

4. The Future

There are three major areas to future development.

A. Refining what already exists using the feedback from the indexers.

B. Enhancing the system with features that make it easier to use. A few ideas that are being explored are:

1) There is a problem in using a 24 line by 80 character screen in that not enough information can be displayed at once. An indexer would like to be able to see the abstract, or title, for instance, while working on the keyword field. Since a DEC VT100 terminal with split screen capabilities is being used, the possibility of scrolling abstract, or title, in the top few lines, while allowing editing of the keyword field on the remainder of the screen will be tested.

2) Currently, keywords can only be entered when an indexer is in the keyword mode in the editor. If a potential keyword appears in the abstract, or as a related term, for instance, in the thesaurus display, the indexer

must return to the menu and then choose the keyword display before entering a term in the keyword field. If a light pen (touch screen, "mouse", etc.) were added, then it would be possible to "point at" any word on the screen and enter it immediately into the keyword field, saving many keystrokes.

C. Finally, of course, there is a wealth of research potential given this basic system. A few thoughts are:

1) Moving towards automatic indexing. The computer, through semantic and syntactic analysis of text could be programmed to "predict" keywords and categories for consideration by the indexers. This would also allow a semi-automatic thesaurus enhancement.

2) Explore the "technology for dialogue engineering" as Gaines [11] describes it. He is developing rules for accomplishing "conversational" computing in the most effective manner.

3) There is almost universal agreement in the literature that "good" indexing is that which promotes effective retrieval. There is a chance, with this system, to move beyond the traditional recall, relevance and precision measurements and add a focus on the link between the indexer and user, on the type of communication they can establish through dialogue or files. Retrieval strategies on the part of the user could be fed back to the indexer and the indexer should be able to make use of previously indexed documents in the database. Walker [12] is prototyping a system where the major objective is to gather data about the nature of problem formation and then is modifying the ways in which material is organized in the file and presented to the user. The "user", in our case, is also the "indexer". But if a major problem in information science, as Walker states it, is "facilitating of effective communication between human generator and human user" then this concept of indexer as user should help further define and refine the indexing process.

REFERENCES

1. The TIC Interactive System for Processing Energy Information, TID-27598, TIC.

2. DOE/RECON: Manual for the Dial-up User, TID-4758, TIC.

3. Brian W. Kernighan & P. J. Plauger, Software Tools, Addison-Wesley Publishing Co., Reading, Mass. (1976).

4. Jeremy A. Digger, A Study of the Intellectual Elements of Indexing for Information Retrieval. Thesis submitted for fellowship of the Library Association (1973).

5. Guide to Abstracting and Indexing at the Technical Information Center, TID-4583, TIC.

6. DOE Energy Information Data Base: Subject Categories, TID-4584, TIC.

7. DOE Energy Information Data Base: Subject Thesaurus, TID-7000, TIC.

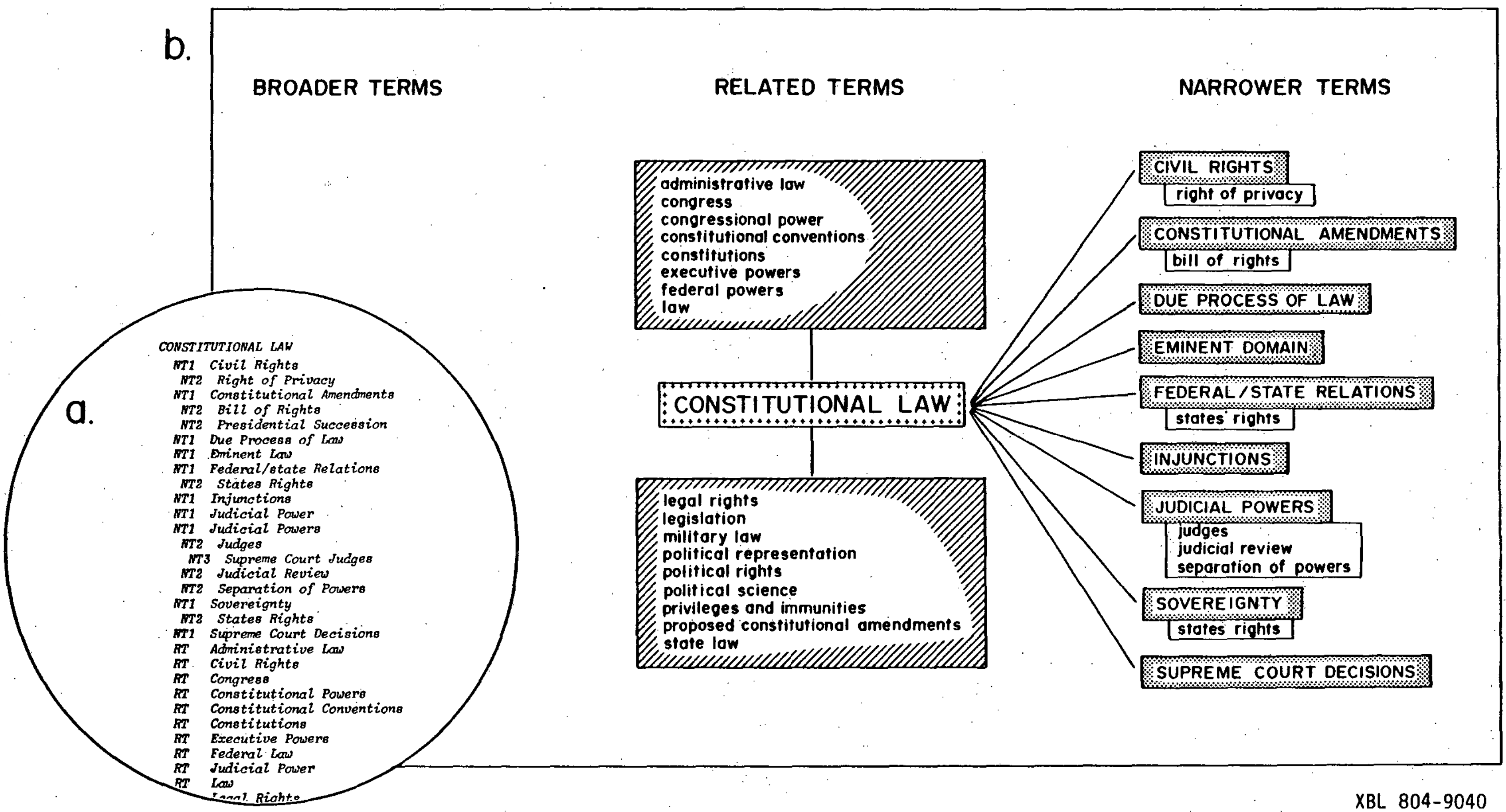
8. D. F. Cahn & C. V. Morano, Interactive Computer Graphics Displays for Hierarchical Data Structures. Lawrence Berkeley Laboratory Report LBL-10247 (May, 1980).

9. Lynn Evans, Evaluation of the ISPRA Automatic Indexing Programs SLC-II. Final Report, INSPEC (July, 1978).

10. Ray W. Graves and Donald P. Helander, A Feasibility Study of Automatic Indexing and Information Retrieval, IEEE Transactions on Engineering Writing and Speech, Vol. EWS-13(2), pp. 58-59 (1970).

11. Brian R. Gaines, The Technology of Interaction-Dialogue Programming Rules, International Journal of Man-Machine Studies, Vol. 14, pp. 133-150 (1981).

12. D. E. Walker, The Organization and Use of Information: Contribution of Information Sciences, Computational Linguistics and Artificial Intelligence, JASIS, Vol. 32(5), pp. 347-363 (1981).



XBL 804-9040

Figure 1: Typical thesaurus word-blocks in traditional (a) tabular-alphabetic and (b) computer graphic presentations. The main-term (MT) appears at screen center, broader terms up to two levels distant (BT1) (BT2) in the left column, narrower terms (NT1) (NT2) in the right column, and related terms (RT) in column with the main term.

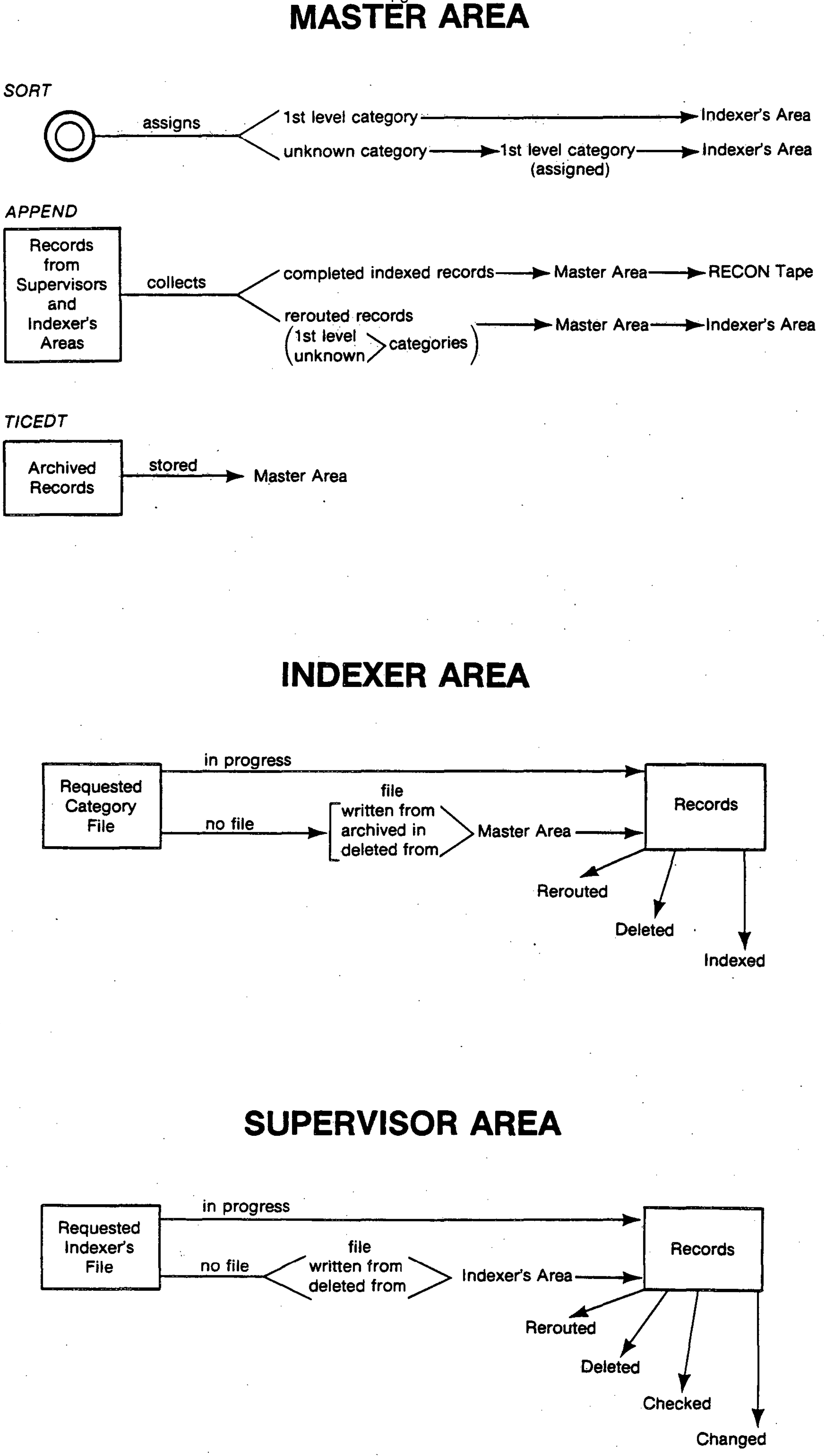

MASTER AREA
SORT
assigns
1st level category
Indexer's Area
unknown category
1st level category (assigned)
Indexer's Area
APPEND
Records from Supervisors and Indexer's Areas
collects
completed indexed records
Master Area
RECON Tape
rerouted records
(1st level / unknown > categories)
Master Area
Indexer's Area
TICEDT
Archived Records
stored
Master Area
INDEXER AREA
Requested Category File
in progress
no file
file written from / archived in / deleted from
Master Area
Records
Rerouted
Deleted
Indexed
SUPERVISOR AREA
Requested Indexer's File
in progress
no file
file written from / deleted from
Indexer's Area
Records
Rerouted
Deleted
Checked
Changed

Figure 2

This report was done with support from the Department of Energy. Any conclusions or opinions expressed in this report represent solely those of the author(s) and not necessarily those of The Regents of the University of California, the Lawrence Berkeley Laboratory or the Department of Energy.

Reference to a company or product name does not imply approval or recommendation of the product by the University of California or the U.S. Department of Energy to the exclusion of others that may be suitable.

TECHNICAL INFORMATION DEPARTMENT
LAWRENCE BERKELEY LABORATORY
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720