Title

USE OF DATA-BASE ACCESS FOR INTERINDEXER COMMUNICATION AND FOR INDEXER TRAINING

Author

Herr, J. Joanne.

Publication Date

1970-07-01

For the 33rd Annual Meeting of the
American Society for Information
Science, October 11-15, 1970,
Philadelphia, Pa.

UCRL-19862
Preprint

c.2

USE OF DATA-BASE ACCESS FOR INTERINDEXER COMMUNICATION AND FOR INDEXER TRAINING

J. Joanne Herr

July 1970

AEC Contract No. W-7405-eng-48

LAWRENCE RADIATION LABORATORY
UNIVERSITY of CALIFORNIA BERKELEY

DISCLAIMER

USE OF DATA-BASE ACCESS FOR INTERINDEXER COMMUNICATION AND FOR INDEXER TRAINING

J. Joanne Herr
Lawrence Radiation Laboratory
University of California, Berkeley, California 94720

Abstract

Data-base access--the availability of all of the indexing for a data base--is discussed as a technique for training indexers and as a method for increasing indexing consistency. Results are presented of an exploratory test of the effect of data-base access on indexing by experienced indexers in a multidisciplinary service, Nuclear Science Abstracts. Indexing style conventions for subject-heading indexing (subject headings with modifier lines are used for the printed index to Nuclear Science Abtracts) were readily recognized by the test participants. The major effect of consulting a reasonably self-consistent data base on descriptor indexing (used for machine searching) was that specific coordinate indexing terms were more available to the participants.

Introduction

At the Lawrence Radiation Laboratory—Berkeley we have been studying the feasibility of decentralized indexing for Nuclear Science Abstracts, which is currently produced at the AEC's Division of Technical Information Extension (DTIE), Oak Ridge, Tennessee. In addition to feasibility under conditions currently in force at DTIE (these are outlined in Reference 1), we are interested in techniques to compensate for the communication problems brought about by decentralization.

Data-base access is commonly employed by catalogers when they consult the catalog to determine how material similar to an item to be entered into the collection has already been cataloged. Similarly, indexers often compare a new document with items under various subject headings to determine the most appropriate slot for the new document.

Full use of the method of data-base access involves consulting an index to find actual items --with a display of their full indexing--that are similar in content to the new document to be entered into the system. The goal is to produce indexing for the new item that reflects its relationship (similarities and differences) to the rest of the collection. A deliberate approach to data-base access for indexers through an on-line system has been discussed by Bennett (2).

Data-base access seemed to be a method by which decentralized indexers could communicate through their work. By having the full indexing for the collection--or a large enough portion of it--new indexers could be trained with minimal contact with experienced indexers by attempting to duplicate indexing patterns in the system. An exploratory study of the use of the technique for training indexers and of its effects on interindexer consistency is presented here.

Construction of the LRL File

The preliminary work on decentralized indexing at LRL was an investigation of subject-specialized indexing. This is the technique used at DTIE to produce the indexing for Nuclear Science Abstracts. Indexing for the Atomic and Molecular Physics subsection of NSA was carried out at LRL in parallel with the indexing for that subsection at DTIE for approximately 5 months. This measurement phase was preceded by about 2 months of training by correspondence between the subject-specialist indexer at DTIE assigned to the subsection and the LRL indexer.

An inverted file, stored on Port-A-Punch cards, was constructed by the LRL indexer during the training phase of the decentralized indexing project. In the linear file were the indexing worksheets--complete with abstracts, descriptive cataloging, and subject-heading and descriptor indexing. (Subject headings with modifier lines make up the printed index to NSA; descriptor indexing is based on the EURATOM Thesaurus and is used for the machine-searchable coordinate index.) Although the inverted file could have been based on subject-heading indexing, it was felt that selector indexing would be more useful because it provided more entry points to the linear file.

During the training phase, the file was used to locate items previously indexed at LRL as well as comments and corrections from the indexing instructor at DTIE. After the training period, parallel indexing was carried out at LRL and DTIE. The file was kept up during this phase of the project, primarily as a means for enhancing the LRL self-consistency.

Test of Usefulness for Interindexer Consistency

A test of the file's usefulness to experienced indexers was carried out at DTIE. (Details of the test are discussed in Reference 3.) At the time of the test, the linear file contained 205 items and the inverted file, 470 descriptors; this represented about 3 months of indexing for the Atomic and Molecular Physics subsection of NSA.

A group of six indexers, all from Physics or Chemistry sections at DTIE, participated in the test. Two of the six occasionally indexed material for the subsection for which the file was constructed. The General Physics section chief chose two sets of ten papers each as representative of material in the subsection.

Each participant first indexed one of the sets of ten papers using only the tools normally employed--the subject-heading authority list, the NSA index, and the EURATOM Thesaurus. Each was then given a copy of the inverted and linear files, and he revised his earlier indexing to conform, as closely as possible, with the style found in the LRL file. Some examples of searching the file for precedents were explained before the participants used the file. Comments on its use, including inability to find precedents, were requested from the participants.

Six of the ten papers in each set were current and were indexed by the indexing instructor and by the LRL indexer. The indexing of these six papers by the participants was compared with that by the instructor, whose work was used as the standard because the file was based on his indexing style. The comparison of the LRL indexing with that by the DTIE instructor (the standard) indicates the norm for two indexers who have worked fairly closely for an extended period.

Results

No significant effect of the LRL file on the depth of indexing as measured by the number of entries was found. Following use of the file, however, a shift was observed to more specific descriptors, which are valuable for retrieval. Apparently, not all the more specific terms for this subsection were originally a part of the participants' working vocabularies. In some cases combinations of general terms (e.g., ATOMS + BEAMS) were replaced by an appropriate specific term (ATOMIC BEAMS); in other cases, more specific terms were simply added to enhance the specificity of the indexing. These shifts to more specific terms tended to increase the consistency between the participant and the standard (the indexing instructor). The total number of such changes was not large but each of the six participants made at lease one.

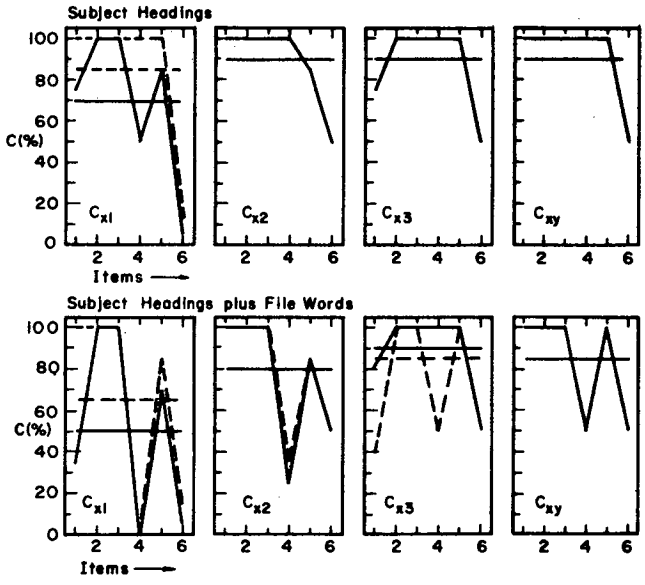

Figure 1. Consistencies, C_{x1}, C_{x2}, C_{x3}, and C_{xy} of participants A1, A2, and A3, and the LRL indexer (y) with respect to the reference indexer (x) for subject-heading indexing.

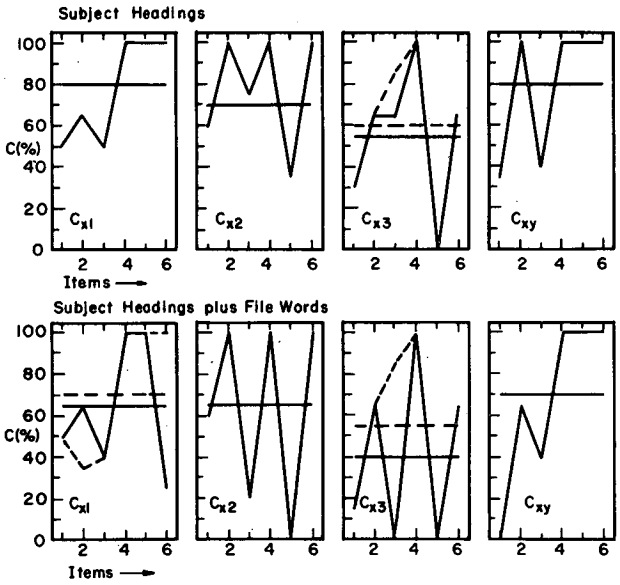

Figure 2. Consistencies, C_{x1}, C_{x2}, C_{x3}, and C_{xy} of participants B1, B2, and B3, and the LRL indexer (y) with respect to the reference indexer (x) for subject-heading indexing.

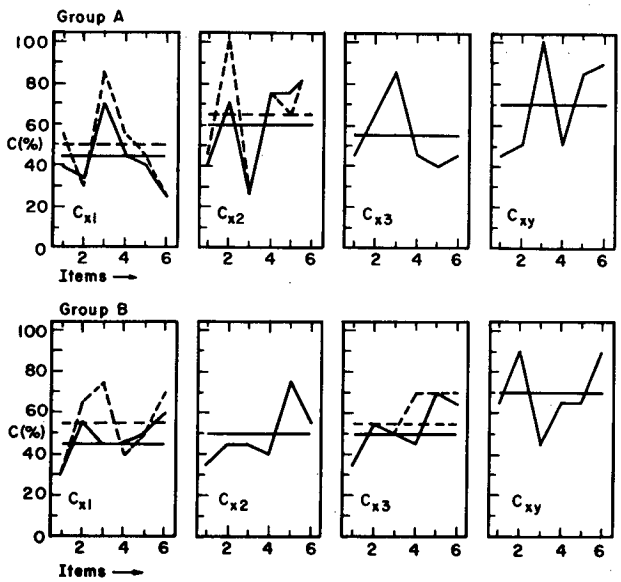

Figure 3. Consistencies of the participants, A1-A3 and B1-B3, and the LRL indexer (y) with respect to the reference indexer (x) for descriptor indexing.

Figures 1, 2, and 3 illustrate the item-by-item consistencies of the participants (A1, A2, A3, B1, B2, and B3) and of the LRL indexer (y) with respect to the reference (x) before (solid lines) and after (broken lines) use of the file. The overall averages before and after use of the file are given by solid and broken horizontal lines, respectively. Consistency for a pair was calculated as the ratio of the number of terms in common to the total number of unique terms of the pair (4). Figures 1 and 2 show the results for subject headings and subject headings plus file words (first word in the modifier line); descriptor-indexing results are given in Fig. 3.

These figures indicate that the effect of the file depended strongly on the individual participants and on the paper being indexed. The dependence on the paper stems from the distribution of subject matter in the file and its limited size. There were not precedents for all the items in the test. For instance, Fig. 2 shows that participant B3 changed his subject-heading indexing for only one (number 3) of the six items. Two others that appear to be candidates for change (numbers 1 and 5) were not well precedented in the file; it would be unreasonable to expect the file to have any effect on the indexing for these items.

If a precedent existed, most of the participants were able to find it. Unfamiliarity with the descriptor vocabulary occasionally made entry into the file difficult. The participants felt able to detect indexing conventions, but their indexing indicated that conventions in subject-heading indexing were easier to detect (and implement) than were descriptor conventions. This is not unreasonable in light of the participants' experience in using the NSA printed index, the smaller number of types of subject-heading entries, and the phrase-like nature of the modifier lines.

Most of the increases in subject-heading consistency were due to recognition of indexing conventions by the participants; the strongest effect of the use of the file on coordinate indexing was an increase in the specificity of the descriptors used.

In every case but one in which changes were made, the average consistency did increase, although the consistencies for individual items did not always increase. The exception indicates a problem inherent in basing new indexing on that done previously if the earlier work is not itself completely consistent. Participant A3 changed his subject-heading indexing to agree with one early item in the file that was not consistent with the bulk of the file. In a data base prepared by more than two indexers the number of inconsistencies would surely be even larger than was the case with the LRL file. However, detection of inconsistencies through searching the data base should lead to resolution of the differences in approach.

The participants found the Port-A-Punch cards inconvenient to use. One of them referred to the problems of "card shuffling." A printed (or book) index would have been preferred. The on-line system described by Bennett would probably be as effective as a printed index, or more so. The participants felt that an indexer's guide is a more concise and explicit way of controlling the handling of common types of material, but that access to the data base would be useful for unusual items.

An interesting effect on selector-indexing vocabularies was observed. In a multidisciplinary system, a "controlled" vocabulary is controlled in reality only to the extent that the specific terms needed are actually available to the indexers. That the working vocabulary of an indexer should not include the whole thesaurus [for instance, the EURATOM Thesaurus contains 13,000 valid terms (5)] is not unreasonable. Nor is it unlikely that an indexer should have difficulty in finding in the sea of terms those that he needs. Entrance into the data base permits examination of the vocabularies of other indexers and provides a method for increasing an indexer's working vocabulary. Another technique for making the appropriate terms available is to increase their visibility by separating a cluster of related terms from the great bulk of the tesaurus. This is done, for instance, in the MEDLARS thesaurus and, for a part of the vocabulary, by the EURATOM Terminology Displays. An approach that we have taken (6) is the arrangement in clusters of the set of core-concept terms needed for indexing a subsection of Nuclear Science Abstracts (generally the level of responsibility of a subject specialist at DTIE). Current studies at LRL indicate that the number of terms needed by a subject-specialist indexer to cover 90% or more of the input for his subject is of the order of 600.

Summary

The results of a limited test of the usefulness to experienced indexers of access to a data base have been presented. Because the technique was new to the participants and because so few papers were indexed, this test can be considered to be only an exploratory quantitative study. Of great importance to this evaluation were the comments of the experienced indexers, from which were gleaned indications of effectiveness, suggestions for modifications, and possible alternatives.

In most cases, the use of the LRL file did result in increased consistency between the participants and the reference indexer, although not all the participants used the file to the same degree. As the participants were all experienced indexers, the lack of startling changes in their indexing is not unexpected. The greatest potential value of data-base access is for establishing indexing conventions and detecting inconsistencies and as a means of training indexers through concrete examples.

The use of the LRL file during the training phase of the decentralized indexing project did indicate that such a file was a convenient and efficient method for storing and retrieving examples of indexing and notes on indexing. It appears that a trainee's attempts to reproduce indexing of an experienced indexer can to some extent replace item-by-item correction by an instructor.

References

1. Robert L. Shannon, "Nuclear Science Abstracts: A 21-Year Perspective," presented at the Symposium on the Handling of Nuclear Information, Vienna, Feb. 16-20, 1970, IAEA/SM-128/32.
2. John L. Bennett, "On-Line Access to Information: NSF as an Aid to the Indexer/Cataloger," American Documentation 20 (3), 213-220 (1969).

3. J. Joanne Herr, "Effects of Data-Base Access on Indexer Consistency", Lawrence Radiation Laboratory Report UCRL-19250, July 1970.
4. R. S. Hooper, "Indexer Consistency Tests --Origin, Measurements, Results, and Utilization," presented at the 1965 Congress International Federation for Documentation, Washington, D.C., October 10-15, 1965.
5. R. Colbach, "Thesaurus Structure and Generic Posting," presented at the Symposium on the Handling of Nuclear Information, Vienna, Austria, February 16-20, 1970, IAEA-SM-128/33.
6. G. L. Smith, J. J. Herr, and R. K. Wakerling, "An SDI System Based on NSA Magnetic Tapes: User Profiling and the Implications of Decentralized Indexing" Lawrence Radiation Laboratory Report UCRL-19290, Dec. 1969, same symposium, IAEA-SM-128/33.

TECHNICAL INFORMATION DIVISION
LAWRENCE RADIATION LABORATORY
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720