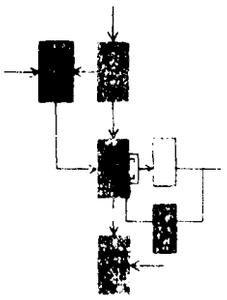
August, 1968

REPORT ESL-R-360
M.I.T. PROJECT DSR 70054
Research Grant NSFG-472 (Part)



GENERATION AND ENCODING OF THE PROJECT INTREX AUGMENTED CATALOG DATA BASE

Alan R. Benenfeld

Electronic Systems Laboratory

Project Intrex Group

MASSACHUSETTS INSTITUTE OF TECHNOLOGY, CAMBRIDGE, MASSACHUSETTS 02139

Department of Electrical Engineering

August, 1968 ESL-R-360 Copy_____

GENERATION AND ENCODING OF THE PROJECT INTREX AUGMENTED CATALOG DATA BASE

by

Alan R. Benenfeld

The research reported in this document was made possible through the support extended the Massachusetts Institute of Technology, Project Intrex, under a grant from the Carnegie Corporation, and under Research Grant NSF-C472 (Part) from the National Science Foundation and the Advanced Research Projects Agency of the Department of Defense.

Catalog Data Input Group
Electronic Systems Laboratory
Department of Electrical Engineering
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

ABSTRACT

This report is based upon a paper presented to the 6th Annual Clinic on Library Applications of Data Processing, University of Illinois, Urbana, Illinois, on May 7, 1968, and it will also appear in the published proceedings of that conference.

A flexible, analytically-structured, catalog-record format was designed to aid in meeting the objectives of the display-oriented Project Intrex augmented catalog experiments. The analytical format, and the catalog data elements and their encoding for machine readability are discussed. The selection of documents from the literature of materials science and engineering for the Intrex data base, the generation of catalog records of those documents, and the initial processing of those records for computer-storage are covered. Initial studies that were made to evaluate the processing of catalog records receive attention. One study shows that data input at an on-line terminal in our current MIT CTSS operating environment is twice as expensive as our normal off-line data input using punched paper tape. Attention is also given to the creation from each document of a set of complete index term phrases and to the problems of matching these unconstrained terms with similarly unconstrained subject request phrases. Computer programs for phrase decomposition and word stemming, and interactive man-machine dialog, will help solve the problems of subject retrieval. The main development phase of the experimental time-shared augmented catalog is nearing completion.

ACKNOWLEDGEMENT

I want to express my gratitude to the members of the Project's data input and storage and retrieval groups for their advice and comment. In particular, my thanks are extended to Mrs. Jane Marshall, Mrs. Betty Gurley, Professor J. F. Reintjes, and Messrs. Peter Kugel, Robert Kusik, and Richard Marcus.

This report is based upon a paper presented at the 6th Annual Clinic on Library Applications of Data Processing, University of Illinois, Urbana, Illinois, May 7, 1968; it will also appear in the published proceedings of the conference.

CONTENTS

LIST OF FIGURES

LIST OF FIGURES (Contd.)

I. INTRODUCTION

Project Intrex is a series of long-range experiments in information transfer which is meant to have particular application to university library services in the next decade. The experimental work currently under active investigation incorporates several of the ideas generated during the summer of 1965 at the Intrex planning conference.[1*] Essentially, this research falls into two major areas: (1) an augmented catalog, which stores and retrieves information about documents, and (2) text access, which stores and retrieves documents.[2]

Work on the augmented catalog is divided among a data input group, a storage and retrieval group, and an output system (display console) group working together with the text access group. This paper will primarily cover those aspects in the development of the augmented catalog experiments concerning data input, namely, the generation and encoding of the catalog's data base. In order to appreciate the data structure, we must first consider the nature of the catalog experiments.

The objectives in the design of the Intrex augmented library catalog are to provide answers to such questions as:

> What data elements in a catalog are necessary and useful to different categories of university community users?
>
> What is the best way to encode and to display this data?
>
> What methodologies and searching strategies do people employ in using a catalog?
>
> What costs are involved in supplying various kinds of information?
>
> How can these findings be optimized in the development of a prototype operational catalog?

---

* Superscripts refer to numbered items in List of References at the end of this document.

There are three key terms to keep in mind about this catalog:

1. It is experimental.
2. It is time-shared.
3. It is augmented.

Augmentation refers to the extension in coverage and depth of information in the catalog beyond that of traditional catalogs and indexes. This single catalog accommodates the description of all types of documents, monograph and serial, whole and analytic, print and non-print. Consequently, a large number of bibliographic data elements associated with diverse bibliographic forms have been identified and synthesized into a hospitable, single-catalog structure.

Time shared means that the catalog is computer-based and is accessible simultaneously to a number of users at remote locations. More significantly, it implies that the catalog is dynamic in nature. From the user's point of view, it means that he engages in dialog with the catalog, he maintains selectivity over the amount of information to be called up from any catalog record, and he even can make entries and comments into the catalog. These interaction capabilities also apply to the librarian as a user, and to them we can add ease of updating and revision whereby such maintenance does not disrupt other users; they also indicate an extension to the concept of access to an evolving catalog entry when a record is created in stages.

Experimental is perhaps the most important key term because it gives us flexibility to investigate the optimum structure of a catalog. There are no ties to existing or past catalog structures which might otherwise constrain achievement of our experimental objectives. There is freedom to change both order and format of data elements to meet varying experimental conditions. While these freedoms exist, compatability is also important and weight is given to the more traditional representations of data elements.

With our experimental flexibility, the augmented catalog is an evolving system whose initial and final configurations can be very different. With the augmentation feature, we build upon the basic records found in traditional catalogs. With the computer base, we have a dynamic system whose data elements can be stored, manipulated, and retrieved to suit varying user needs.

The discussion so far has been intended to create a general flavor for the basis of the augmented catalog. The body of this paper discusses specifics in the current ongoing development of the catalog; actual testing of the catalog in a real-user environment will not begin until the summer of 1968. Encoding of the data base will be discussed first, followed by the generation of the data base, and finally by some initial evaluations.

II. ENCODING OF THE DATA BASE

A. DATA ELEMENTS

The data structure of the augmented catalog was established by identifying, describing, and analyzing data elements associated with diverse bibliographic forms. Those data elements that were the same or similar in purpose were synthesized into one. The aim was to establish consistency in handling the same information from different bibliographic forms, an important factor for this single catalog. Data elements normally implicit in traditional records by form, language, or position in the record, were made explicit. Specifications were written[3] which defined each data element and established its identification (tagging) and format for machine readability. The initial specifications for Project MARC at the Library of Congress,[4,5] and the descriptive cataloging manual of the Atomic Energy Commission's Division of Technical Information[6] were particularly helpful to us.

There are approximately 115 data elements that we have chosen to work with to date. These data elements are combined into 50 fields. Each data element is equivalent to a subfield. Each field is assigned a number. The 50 fields can be grouped into six major categories:

I. Catalog control information (fields 1-5).
II. Physical document control information (fields 10-12).
III. Descriptive cataloging information (fields 20-50).
IV. Document content information (fields 65-73).
V. Article citation information (field 80).
VI. User feedback information (field 85).

The data elements are listed in Fig. 1 under their assigned field numbers.

Naturally, only some of the data elements will apply to the description of any one record. It is interesting to note that Curran and Avram[7] have compiled a listing of several hundred bibliographically-related elements; yet any existing traditional catalog or index contains only a small fraction of these elements.

It is not possible to cover in this paper the meaning and specifications for all data fields that can constitute a catalog record. The discussion, together with the illustrations, is intended to indicate the overall nature and structure of the viable total record and to show the relations between parts of a record.

B. RECORD AND FIELD IDENTIFICATION

Each catalog record begins with the tag for field 1, the record number. This tag, in machine-readable form, identifies field 1 to the computer; and because of its initial position, it also serves to identify to the computer the beginning of a new catalog record. Our current initial mode of machine-readable input is punched paper tape. The field 1 tag is encoded on this tape as //1/ where the two initial slash marks are preceded by the nonprinting punch codes for carriage return and lower case. Other fields are similarly identified by their field numbers. However, the order of appearance of other fields in a record may be random. The end of a catalog record is machine identified by codes for the 3-character string -.-, that is, the 3 characters "hyphen period hyphen".

C. CODING

The data entered into a field may be natural data, or fully or partially coded data. Data elements that are coded, with some examples, are indicated in Fig. 1. Some codes are mnemonic, particularly where the number of individual items to be coded in our catalog is small, as are library names, or where such a device would facilitate the cataloging operation, as with relators. These codes, like the rest of the catalog format, can be expanded or changed to meet new requirements. Only eight fields are completely coded; these are marked by an asterisk in Fig. 1.

The rest of this discussion on encoding is aided by reference to Figs. 2 and 3, both of which show a complete catalog record, and to Figs. 4 and 5, which give additional examples of specific formats of some fields. The names of the fields appearing in the record illustrations can be obtained through the field numbers from the listing in Fig. 1.

AUGMENTED CATALOG DATA ELEMENTS

I. CATALOG CONTROL INFORMATION

Field Number and Name	Data Elements
1. Record Number	Record Number
2. *Document Selection	Selector's Position Code (1 = librarian) (2 = faculty member)
	Research Group Code (B = high temperature metallurgy)
3. Input Control	Worker Code Number (1 = Jane Rust)
	Task Code Number (2 = subject cataloging)
	Date Task Done
	Time to Do Task
4. On-Line Date	On-Line Date
5. Access Number	Microfiche Number
	Inclusive Frame Numbers

II. PHYSICAL DOCUMENT CONTROL INFORMATION

10. L. C. Card Number	L. C. Card Number
11. Library Location and Number of Copies	Library Name Code (e = Engineering Library) (m = Materials Center Reading Room)
	Number of Copies
	Call Number
	Special Shelf Location Code (E = in journal stacks) (M = on reference shelf)
	Non-Call Number Shelf Arrangement Code (1 = under title) (2 = under course number)
12. Serial Holdings	Library Name Code (see field 11)
	Holdings

Fig. 1 Fields and Data Elements in the Augmented Catalog. Parentheses enclose examples of coded data elements. Asterisks indicate completely coded fields.

III. DESCRIPTIVE CATALOGING INFORMATION

Field Number and Name	Data Elements
20. *Main Entry Pointer	Main Entry Pointer (1 = personal name) (3 = title)
21. Personal Names	Personal Name
	Honorific
	Dates
	Personal Name Relator Code (ED = editor) (TS = thesis supervisor)
22. Personal Name Affiliations	Title of Position
	Corporate Name
	Place Qualification
	New or Former Affiliation
23. Corporate Names	Corporate Name
	Place Qualification
	Non-Place Qualification
	Conference Name
	Corporate Name Relator Code (TH = thesis submitted to this institution) (HA = institution at which a conference was held)
24. Title	Main Title
	Supplied Title
	Subtitle
	Translated Title
25. Coden Title	Coden Title
26. Edition Statement	Edition Statement
27. Publisher	Publisher or Distributor Name
	Non-Publisher Relator Code (DS = distributor) (IA = issuing agency)
28. Place of Publication	Place of Publication
29. Dates of Publication (for monographs and monographic sets)	Publishing Date
	Copyright Date
	Reproduction Date

Fig. 1 Continued

Field Number and Name	Data Elements
30. *Medium (physical nature of the document)	Medium Code (1 = conventionally printed) (6 = microfiche) (18 = constructional model)
31. *Format (arrangement of information within a document)	Format Code (f = directory) (r = professional journal) (bb = article like that found in a professional journal) (ii = editorial)
32. Pagination	Pagination
33. Illustrations	Illustration Note
34. Dimensions	Dimensions
35. Serial Frequency	Issue Frequency Code (a = daily) (g = bimonthly)
	Number of Issues Per Serial Volume
	Number of Volumes Per Year
36. *Language of Document	Language Code (e = English) (f = French) (r = Russian)
37. *Language of Accompanying Abstract	Language Code (see field 36)
38. Series Statement	Series Name, Including Title
	Number in the Series
39. Report Numbers and Patent Numbers	Report Number
	Paper Number
	Patent Number
	Patent Country of Origin
	Based-upon Relator Code (BR = based on a report bearing the accompanying number)
40. Contract Statement	Contracting Agency Name or Contract Monitor Name
	Contract Number
41. Supplement Referral	Type of Supplement Code (st = supplement to) (ib = indexed by)
	Number or Date of Supplement
	Record Number of Supplement

Fig. 1 Continued

Field Number and Name	Data Elements
42. Errata	Location of Incorrect Data
	Corrected Data
	Corrector's Name
	Citation to Published Errata
43. Thesis	Degree Level Code (1 = doctorate)
	Date of Thesis
	Degree Abbreviation
	Subject Field of Degree
44. Variants	Language Code for the Variant (see field 36)
	Medium Code for the Variant (see field 30)
	Library Location of Variant (see field 11)
	Source Document
45. Titles of Variants	Titles of Variants
46. Article Receipt Date	Date Article Received for Publication
47. Analytical Citation Statement	Serial Analytics: Coden Title
	Numbers - Volume, Issue, Part
	Date
	Inclusive Paging
	Monograph Analytics: Page or Chapter Location
	Transfer Code to Record of Larger Work
48. Abstract Services	Coden for Abstracting Publications
49. Cost - Text Access Source	Price
	Medium Code (see field 30)
50. Commercial Cost and Availability	Price
	Medium or Format Code (see field 30 and field 31)
	Language Code (see field 36)
	Supplier Name and Address

Fig. 1 Continued

IV. SUBJECT CONTENT INFORMATION

Field Number and Name	Data Elements
65. *Author's Purpose	Author's Purpose Code (t = report on original research -- theoretical) (e = report on original research -- experimental) (n = review--non-critical)
66. *Level of Approach	Level of Approach Code (1 = professional (including graduate level) (4 = undergraduate level)
67. Table of Contents	Heading
	Beginning Page Number
68. Special Features	Special Features Statement
69. Bibliography	Type of Reference (1 = references) (2 = suggested readings)
	Location of References (e = end of complete text) (f = footnotes)
	Number of References
70. Excerpts	Excerpt
	Location of the Excerpt
71. Abstracts	Abstract
	Abstractor
	Citation to the Abstract
72. Reviews	Review
	Reviewer
	Citation to the Review
73. Subject Indexing	Subject Term
	Weight (1 = term representing most of document content) (2 = term representing major section of document content) (3 = term representing small segment of document content) (4 = term representing materials, tools, techniques not appearing in another index term) (0 = term generic to document content)

Fig. 1 Continued

V. ARTICLE CITATION INFORMATION

Field Number and Name	Data Elements
80. Article Reference Citations	References Cited
	Citing References

VI. USER FEEDBACK INFORMATION

85. User Comments	User Comments

Fig. 1 Continued

//1/ 4299

//2/ A24

//5/ 7-C10-D3

//20/ 1

//30/ 1

//69/ 1e(4)

//33/ illus.

//36/ e

//37/ f

//65/ b

//66 3

//31/ dd

//47/ pp.1453-1458. IN &+4180&

//24/ An internally reflecting optical resonator with confocal properties

//21/ Holshouser, D.F.

//22/ University of `Illinois', `Urbana'. Electrical Engineering Dept.

//40/ `U.S.' Air Force Office of Scientific Research #AF-49-(638)-556#AFOSR-62-250

//66/ Contains diagram of geometry for confocal internal reflection

Fig. 2 Sample Catalog Record for a Conference Paper. Underlined field numbers indicate fields formatted by librarians. Dots indicate fields tagged for the typist's attention.

//70/ Optical resonators using spherical mirrors, e.g. confocal systems, have been shown to have significant advantages over configurations using planar mirrors. In particular, diffraction losses can be much lower and alignment is less critical. However, planar systems have had an advantage heretofore in that coated mirrors could be replaced by internally reflecting prisms, thereby eliminating the problems associated with lossy or fragile coatings. Also, undesired modes are reduced since rays not parallel to the axis are not completely reflected. This paper describes the configuration for an internally reflecting surface which exhibits properties of a spherical mirror, and presents experimental results obtained with a semi-confocal maser using this configuration. [text, p.1453]

//73/ internally reflecting optical resonator with confocal properties (1);
configuration for an internally reflecting surface which exhibits properties of a spherical mirror confocal system (1);
basic properties of a confocal resonator (2);
analytic expression for the internally reflecting surface which satisfies confocal requirements (3);
Schott barium crown glass doped with neodymium (4);
fabrication of semi-confocal optical maser (3);

//3/ 37/2, 032968, 11:15-11:25;
1/1, 040168, 9:26-9:29;
1/7, 040268, 11:06-11.08;
11/4, 040468, 11:35-11:50;
-.-

Fig. 2 Continued

//1/ 3644

//2/ A24

//5/ 175-A1-B5

//20/ 1

//30/ 1

//47/ PHRVA. v.161,no.2,091067. pp.350-366.

//21/ Hempstead, Robert D. (TA);
Lax, Melvin (JA)

//22/ Bell Telephone Laboratories, 'Murray Hill', 'N.J.'/
University of 'Illinois', 'Urbana'. Dept. of Physics;
Bell Telephone Laboratories, 'Murray Hill', 'N.J.'

//23/ M.I.T., 'Cambridge', 'Mass.' (BT);
#American Physical Society Meeting, 'New York', 1966

//24/ Classical noise, part 4; noise in self-sustained oscillators near threshold

//43/ 2,09006S. M.S. (Electrical Engineering)

//31/ bb

//46/ 032867

//36/ e

//37/ e

//33/ illus.

//69/ 1f(38)

//67/ 1. Introduction (p.350)
2. The Langevin and Fokker-Planck equations (p.352)
3. Transformation to polar coordinates (p.354)
4. Integration over the phase variable (p.354)
5. Steady-state amplitude probability distribution (p.354)
6. Calculation of the power spectra (p.356)
7. Eigenfunction expansion (p.358)
8. Accuracy of computations (p.362)
9. Summary (p.362)
10. Appendix A. The laser model (p.363)
11. Appendix B. The circuit model (p.364)
12. Appendix C. The boundary condition for G(r, r*sub 0*; *lambda*, *omega*) (p.366)

//63/ Contains tables of fluctuation data and graphs of power spectra and fluctuation potentials

Fig. 3 Sample Catalog Record for a Journal Article. Underlined field numbers indicate fields formatted by librarians. Dots indicate fields tagged for the typist's attention.

//65/ t

//66/ 13

//71/ Because of the relative narrowness of the threshold region, a general model for spectrally pure self-sustained oscillators (both classical and quantum, including gas lasers) can be reduced, in the threshold region, to a rotating-wave Van der Pol (RWVP) oscillator... [Body of abstract omitted from this illustration because of its length]... Thus the intensity fluctuation spectrum is Lorentzian below and well above threshold, but more complex in the threshold region. (author)

//73/ mathematical development of classical noise in self-sustained oscillators near oscillation threshold (1);
Lax-Louisell model for self-sustained oscillator (4);
Lax-Louisell study of laser noise (3);
normalized rotating-wave Van der Pol oscillator (2);
gas laser (4);
laser noise (0);
phase, intensity, and amplitude fluctuations in self-sustained oscillators near threshold (2);
nearly Lorentzian nature of power spectra of noise in self-sustained oscillators near threshold (2);
exact calculation of power spectra in the normalized RWVP oscillator near threshold by numerical Fokker-Planck methods (2);
scaled Langevin equation (4);
Fokker-Planck, Green's function, and eigenfunction methods of calculating power spectra of noise in self-sustained oscillators near threshold (2);
white-noise sources in a self-sustained oscillator (3);
steady-state amplitude probability distribution (3);
one-sided Fourier transform of the spectrum and intensity spectrum of gas lasers (3);
power spectra boundary conditions (3);
equation of motion of a self-sustained oscillator (4);
effect of power output on power spectra of a self-sustained oscillator (3);
effect of net pump rate on operation of a rotating-wave Van der Pol oscillator (2);
sinusoidal power spectrum of a rotating-wave Van der Pol oscillator (2);
linearization methods of calculating power spectra of noise in a self-sustained oscillator outside the threshold region (2);
nonlinear techniques for calculating power spectra of noise in a normalized rotating-wave oscillator (2);

//3/ 10/2, 012968, 1:30-2:25;
1/7, 012968, 2:40-2:56;
10/1, 021468, 1:40-1:50;
5/4, 040168, 11:35-11:50, 1:20-1:38;
-.-

Fig. 3 Continued

PERSONAL NAME AFFILIATIONS (Field 22)

University of `British Columbia', `Vancouver'. Physics Dept./
Uniwersytet Jagiellonski, `Cracow', `Poland'. (ON LEAVE);

:Sloan Research Fellow: Northwestern University, `Evanston', `Ill.'
Materials Research Center/:Research Physicist: Mallory (P.R.) Inc.,
`Burlington', `Mass.';

CORPORATE NAMES (Field 23)

University of `California', `Berkeley'. Electronics Research
Laboratory.

`Swarthmore' College, (`Pa.') Dept. of Physics.

Homer Research Laboratories [`Bethlehem' Steel Co., (`Pa.')]

#Conference on Magnetism and Magnetic Materials, 12th,
`Washington', `D.C.', Nov. 15, 1966.

Fig. 4 Examples of Corporate Name Field and Personal Name Affiliation Field Formats

LIBRARY LOCATION and NUMBER OF COPIES (Field 11)	e,m,QC761.C74.1966; s,h,QC.J84. v.38,no.3 [E]; e, [Fi]; m/1, [Ei];
SERIAL HOLDINGS (Field 12)	(e) v.5:070063 - to date; (m) &- latest two years only &
SUPPLIED TITLE (Field 24)	(Magnetism and magnetic materials); proceedings. (Quantum electronics); selected conference papers.
REPORT and PATENT NUMBERS (Field 39)	CU-3-66-TR-97; UCRL-14467-T (revision) (BR); Pat. U.S., 3250760; Pat. France, AD-82516;
NOTES (as might be used for language, Field 36)	&-Portugese& &-English, Russian, German, French& j&-(captions and similar labels are in English)&

Fig. 5 Additional Examples of Field Formats

D. SUBFIELDS AND DELIMITERS

Machine identification of data elements is necessary for machine searching and retrieval, for transformation of data into a form suitable for display to a user, or for employing different type fonts for different data elements in a printed output. Each separate machine-identifiable data element is also termed a subfield. Delimiters explicitly establish the boundaries of some subfields while other subfields are implicitly delimited by data context. If a data element or a set of data elements can be repeated within its field, then each occurrence of the set is termed a repeating data group (or repeating subfield group). Repeating data groups are also delimited explicitly or implicitly.

The delimiters used for explicit tagging are generally the punctuation symbols, such as colon and semicolon, or the nesting marks, such as parentheses and brackets. Mirror image nesting symbols are particularly useful visual aids in proofreading operations. With the exception of note delimiters (discussed below), delimiters are defined locally, that is, only in terms of individual field specifications. While this is true, the semicolon most often separates repeating data groups. For example, in the personal name field (field 21), there are as many repeating data groups, separated by semicolons, as there are personal names associated with the document. These names are not just those of authors, but of anyone (except publisher) significantly associated with the document. Within each repeating data group, parentheses following a name delimit a mnemonic relator code subfield identifying the person's functional relation to the document, such as joint author (JA), editor (ED), illustrator (IL), or patent assignee (PA).

For each person named in field 21, his corporate affiliation, if known, is given in field 22. The repeating affiliation data groups appear in the same order as the associated personal names in field 21 and they are separated by semicolons. The affiliation data group for each person may itself contain repeating subfield groups; namely,

1. The person's corporate affiliation at the time the work reported on was performed.

2. The person's present corporate affiliation, or the affiliation from which the person is on a leave of absence; the latter is followed by the relator-like phrase (ON LEAVE).

3. The person's title at each of the two affiliations.

Figure 4 illustrates the format of corporate affiliation data groups; a person's title is delimited by colons, and a slash mark separates the two possible title-affiliation subfield groups for each person.

Corporate names associated with a document are placed in field 23 as repeating data groups separated by a semicolon. Each corporate name may have a relator code, such as (TH) for thesis submitted to this institution, and (SP) for sponsor of a conference. Conference names are additionally tagged by a sharp sign (#) preceding the name. The format of a corporate name consists of a main heading which may have one or more subheadings; two spaces separate each subheading. The geographic location pertaining to the name given by the last subheading is added to the entry. This geographic addition is appended to the main heading; an exception occurs when the place is already part of the main heading. All place names, whether they appear within the main heading or as an addition to it, are tagged with a pair of slanted single quotes. Corporate names that are qualified by the addition of a larger associated corporate body name have that addition enclosed within square brackets and this addition may be further qualified by place. Examples of the corporate name structures may be seen in Fig. 4.

The above discussion of the personal name, corporate name, and affiliation fields is illustrative of the use of explicit delimiters. Explicit delimiters are not necessary when (1) there is no need to separate data elements, (2) adjacent data elements are fixed in length or absolute position with respect to each other, or (3) adjacent data elements are of different character content. Data context provides implicit delimitation of information. A field such as language (field 36) does not require in our record explicit delimiters for its data. This is a coded field and 14 languages (those most likely to occur in our experimental environment) have each been assigned a single-letter alphabetic code. A multilingual document may have up to three languages identified in its record as repeating data groups. Under normal circumstances, this field has a fixed maximum length of three characters, each one representing the same category of information; consequently, there is no need for an explicit delimiter.

E. NOTES

In establishing data element formats, we realized that the following conditions on information from a particular record might arise at some future date:

1. A field is fixed in maximum length, but the information cannot fit within the fixed length.
2. A field is coded, but the information cannot be expressed by any of the established codes.
3. A field requires a specified format, but the information cannot be put into such a format.
4. The information can meet the format specifications for a field but it requires further elucidation.

To overcome these conditions, we have postulated a note, appropriately delimited, which may be inserted within any field. The delimiter preceding a note is the character string "ampersand hyphen" (&-) and the delimiter following a note is an "ampersand" (&). Occurrence of these delimiters results in an automatic program override of the normal format of information at this position, and the computer program acceptance of the note as straight text. There are two kinds of notes: those that completely replace a field, and those that clarify information given in a field. The text of all clarification-type notes is enclosed within parentheses so that, on display, it will be distinguished from the information clarified. The specifications for some fields state particular conditions under which use of a note is anticipated. For example, in the language field (field 36) when a document appears in a language not having an assigned code, or when more than three languages are prominent in a document, the field specifications call for the entry as a note of all of the prominent languages in word form. The use of a note is illustrated in Fig. 5.

F. TRANSFER CODE

A transfer code is used whenever information required for a given catalog record is contained in another record. This is especially useful for relating analytics to their respective whole works. For example, library location and full citation information recorded about

an entire conference proceedings need not be recorded again on the separate records for the individual conference papers. A transfer code is used; it contains only the number of the record referred to, which in this case is the record number for the entire conference. The transfer code is delimited initially by the character string "ampersand plus" (&+) and finally by an "ampersand" (&); it is illustrated in field 47 in the record shown in Fig. 2. The information sought from the record referred to in the transfer code will depend upon the particular field or subfield occupied by the transfer code, and by the question asked of the catalog.

G. DATES, DIACRITICAL MARKS, AND ABBREVIATIONS

Dates are entered into the augmented catalog in a six-position configuration. For example, June 3, 1966 is coded as 060366. If two months or years appear in combination, only the first is entered; thus November/December is taken as November and entered as 110000, while 1954/55 would be entered as 000054. The seasons are equated to the months of January, April, July, and October. Dates for other than the twentieth century are entered as a note. Dates appearing in quoted text, or in a qualification to a conference name, are given in conventional style.

All diacritical marks appearing in foreign language text are ignored. The only general abbreviations allowable for machine-readable data input are those listed in the Anglo-American Cataloging Rules; to these we have added Div. for Division, and M.I.T. for Massachusetts Institute of Technology. Specifications for a particular field may indicate locally allowable abbreviations for that field. Abbreviations appearing in text are generally retained.

H. SPECIAL CHARACTERS

The representation of characters and symbols not on our keyboard is of particular importance. These symbols are defined as special characters. Currently they are represented by a word or other symbol equivalent which is bounded by asterisks. Figure 6 lists some of the special characters defined to date. Some symbols have more than one representation depending upon the context in which they are used. The symbol representations are intended to minimize

SPECIAL CHARACTERS

Å = *A*		∇ = *del*	
$\sim$ = *approximately*		>> = *much more than*	
∂ = *d*		$\sum_i$ = *sum over i*	
* = *star*		{ } = *(* *)*	
$\pm$ = *+/-*		$\oplus$ = *+*	

≡ = *triple bond* (Chemistry)
≡ = *is identical to* (Mathematics)

= *yields* (Chemistry)
→ = *approaches* (Mathematics)
= *transition to* (Physics)

$\vec{a}$,... $\bar{a}$,... = *vector*a,... a*bar*,...

$\dot{a}$,... $\ddot{a}$,... = a*dot*,... a*double dot*,...

$\tilde{a}$,... $\hat{a}$,... = a*tilde*,... a*caret*,...

Greek Letters

Spell out in English in appropriate case

$\alpha, \beta, \gamma, \ldots$ = *alpha*, *beta*, *gamma*,...

A, B, $\Gamma, \ldots$ = *ALPHA*, *BETA*, *GAMMA*,...

Germanic Script

$\mathfrak{J}, \mathfrak{K}, \mathfrak{L},$ = *J*, *K*, *L*,

Superscripts and Subscripts

10^{-6}, σ_f = 10*sup -6*, *sigma*sub f*

Fe_2O_3 = Fe*sub 2*O*sub 3*

$\chi = A(T - T_c)^{-\gamma}$ = *chi* = A(T - T*sub c*)*sup -[*gamma*]*

Fig. 6 Examples of Special Characters and Their Representations

problems in human recognition and readability when the representations themselves must appear in outputs that are print-chain or key limited. At the same time, the tagging will permit machine recognition through table look-up for presentation of the actual character on a cathode-ray tube display output.

I. DOCUMENT CONTENT INFORMATION FIELDS

The fields comprising document content information, article citation information, and user feedback information, deserve particular comment.

Author's purpose (field 65) indicates the author's interest in writing the document. This field distinguishes between original research (theoretical, experimental, humanistic), development and application, review (critical, noncritical), comment (critical, noncritical), text, essay, and proposal. Codes for as many categories as are applicable to an individual document are entered into its record.

Level of approach (field 66) indicates the academic level of the author's intended audience, such as professional (including graduate student) level, undergraduate level (including a professional in another nonrelated field), and layman. A user's personal level in understanding documents in a particular subject area is not necessarily the same as his actual academic position.

Table of contents or section headings are given in field 67 whenever these headings provide subject content information. The beginning page numbers of each section are also included.

Features of a document which are particularly noteworthy in adding to the document's usefulness are described in field 68. Such features might include a glossary, indexes, illustrations, symbol definitions, appendices, discussions and exercises. Data for this field is entered as natural text.

Bibliography in field 69 gives a coded description of the type, location, and estimated number, of bibliographic references cited by a document. Three types of bibliography are identified (references, suggested readings, and comprehensive survey) and three locations are identified (end of text, end of chapter, footnotes). In contrast to field 69, which indicates the nature and size of a bibliography, field 80

(article reference citations) contains standardized formats of the actual references cited by journal articles. More will be said about field 80 in the evaluation section.

Excerpts from a document and an abstract of the document may be given in fields 70 and 71, respectively. An excerpt is followed by its text page location and an abstract is followed by its source. For practical reasons, excerpts are given only in the absence of an author abstract accompanying the document. Similarly, excerpts from document reviews, or references to such reviews, can be entered in field 72.

The most important field in a catalog record is subject indexing, field 73. The subject terms describing a document are complete phrases that, while freely chosen, are primarily based on the natural text of a document; no thesaurus or other authority list is used. Each term is assigned a weight of 0, 1, 2, 3, or 4. Regular weights of 1, 2, or 3, reflect the extent to which a document discusses the concept represented by the term. A weight of 1 signifies that the term is descriptive of most of the entire document; a weight of 2 indicates that the term characterizes a section of the document; and a weight of 3 refers to a term that describes a small fraction of the document. The special weight of 4 is assigned to terms representing mathematical tools, instrumental tools, materials, or applications which are cited in the document but which do not appear in any other index term. The special weight of 0 (zero) is reserved for terms that are generic to the subject matter of a document. Each term, followed by parentheses enclosing its assigned weight, constitutes a repeating data group. There is no bound on the number of terms used to describe a document. The subject indexing process will be discussed more fully in the section on data base generation.

J. USER COMMENTS FIELD

A field for an unusual purpose is user comments, field 85. Comments will be sought from users on any aspect of this computer catalog, including the indexing, the records, and the documents these represent. These comments will be specially stored and periodically printed out for verification and editing. Comments falling within the

sphere of a specific field in a given record will be entered into that field directly, as for instance, errata in field 42, or otherwise appended to the field as a note. Those comments expressing a value judgment on a document, or pertaining in general to a record, will be entered in field 85. Comments will be signed, that is, attributed to their source.

K. CATALOG RECORD STRUCTURE

The structure of traditional book cataloging records is based on the three organizing categories--entry (or heading), statement, and note. Curran and Avram[7] discuss the inadequacy of these categories for handling bibliographic data elements for all forms of documents and their records. The augmented catalog record structure is designed along analytical lines to give maximum flexibility in experimentally handling diverse data elements, and to enhance the retrieval and display functions of the Intrex system.

In an analytical record, statements normally found in the body of the descriptive entry of traditional records are broken into component parts, and data of the same kind are listed as a repeating data group. While it is possible to reconstitute traditional statements from listings in an analytical record, this would be inefficient if system output is primarily oriented to producing traditionally formatted printed records. In Intrex, system output is display oriented and the analytically structured record gives added versatility in optimizing displays of bibliographic data.

Still, if full statements are to be generated from an analytical record, the wording and order may not necessarily be the same as appears on a document title page or in a traditional record. These discrepancies are not considered serious for Intrex because the essential value (content or argument) of each element is retained in the analytical record, and because a document title page can be consulted by display through the Intrex text access system.

Our analytical approach can be more formally related to traditional record formats in that statements essentially have been reduced to headings in the analytical record. For example, there is no author statement but, as we have seen, personal and corporate names

associated with the document are listed in an entry-like form. The Anglo-American Cataloging Rules are followed in establishing the form of these entries, with the added provision that corporate names are usually to be qualified by place. A reconstituted author statement would require an appropriate translation of the relator codes associated with a name but this wording may differ from that found on the document's title page. Title and series added entries in traditional records are generally the same as their respective statements; consequently, we can consider title and series statements as headings and their form in the augmented catalog record is not very different from that in the traditional record. The edition statement is also essentially unchanged, but imprint and collation in the augmented catalog record are broken into component fields. An advantage over the paragraphing of traditional records is seen in the treatment of notes. Notes either are tagged and appended directly to the fields to which they apply, or, as in the case of contents, they constitute an individual field in the augmented catalog.

In a retrieval and display-oriented, computer-stored catalog with inverted file directories of names, the concept of a main entry may seem outmoded. Nevertheless, the augmented catalog retains the main entry concept for two related reasons. First, we must still interface with traditional systems. Second, in the absence of standardized bibliographic citations, the main entry aids in answering the user question, "How can I cite this document?" The second reason also has practical application in any limited printed byproducts from the computer-based catalog. Choice of main entry is based upon the Anglo-American Cataloging Rules. A pointer is entered in field 20 indicating whether the main entry is a personal name, corporate name, or title; if the main entry is a name, that name appears first in the appropriate field.

The analytical catalog structure is flexible so that new fields of information can be readily incorporated whenever their desirability is warranted. At this time, specifications are now being written for fields covering the history of a publication title, and for titles other than the main title by which a document may be known.

III. GENERATION OF THE DATA BASE

The processes involved in the generation of a machine-readable data record as input to the augmented catalog are discussed in this section. A general flow diagram of the Intrex system, operating within the M.I.T. Compatible Time Sharing System (CTSS) environment, is shown in Fig. 7. The lower portion of that diagram, the present input work-flow, is shown in more detail in Fig. 8. Each block indicated will be covered in turn; emphasis is on current procedures.

A. DATA BASE AND LITERATURE SELECTION

The literature base for the augmented catalog experiments is in the large interdisciplinary subject of materials science and engineering. Because the current literature for this entire field greatly exceeds the 10,000 document initial size of the first experimental catalog, only literature in selected areas of materials science and engineering is cataloged. These selected areas reflect the research interests of particular groups at M.I.T.; in point of fact, the particular research groups are chosen first. Such selectivity assures us of a specific population of experimental users and it assures the user groups of a meaningful catalog.

Careful attention is given to the identification and selection of research groups. Factors we consider include: size, composition, and stability of the research group; diversity of interests within a group; scientific relationships of a group's interests to those of other groups; and the library and literature orientation of a group.

To develop comparisons on these points, we have held discussions with administrative personnel in the M.I.T. Center for Materials Science and Engineering, and have made appropriate data compilations from two series of annual reports issued by the Center. Additionally, groups were ranked by a number which is the ratio of a weighted group size to the number of different research projects within the group. This factor, together with the compilations and all the comments from

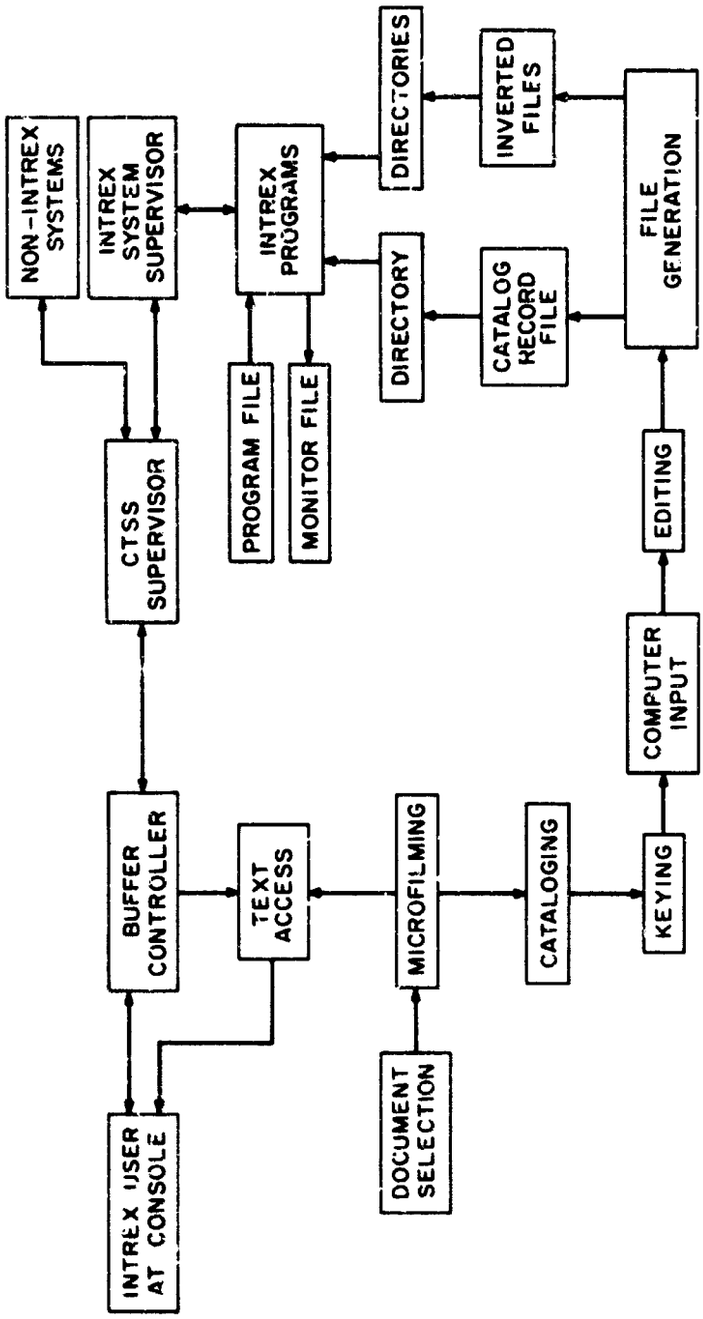


Fig. 7 Generalized Flow Diagram of the Intrex Augmented Catalog - Text Access System in the MIT CTSS Environment. The lower portion emphasizes the data input processes.

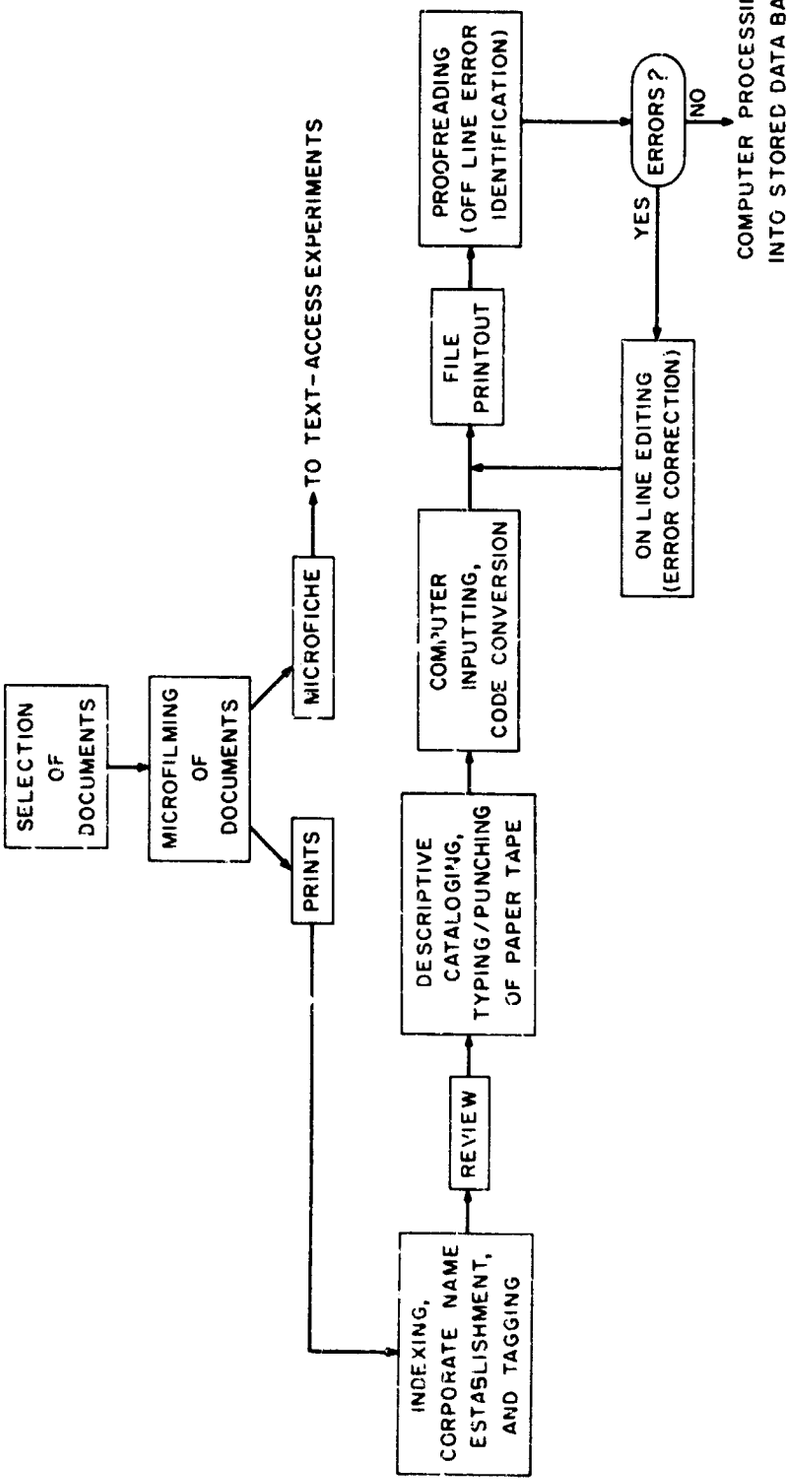


Fig. 8 Current Workflow of the Data Input Processes to the Augmented Catalog

the discussions, led to the initial selection of two research groups. Each group's main interest is considered a subfield of materials science and engineering; these subfields are:

1. Radiofrequency, microwave, and optical spectroscopy of liquids and solids.

2. High-temperature metallurgy.

Additional research groups have since been selected on the basis of modified criteria, namely to increase the number of potential system users and the adequacy of the data base without unduly enlarging that base. As a result, the interests of the first two groups have been considered as dual cores; interests of additional research groups must then significantly overlap with one or the other of the core groups. To date, one group on microwave and quantum magnetics has been added to the first core, and another group on casting and solidification has been added to the second core.

In order to ensure that the literature base we develop will be responsive to the needs of a selected group, and so that group members will be prepared to experiment with it when it is ready, several additional steps are taken. These steps include: eliciting the cooperation of the group; explaining to the group the salient features of the augmented catalog; learning about the group's approaches to the literature; learning which journals are most important to their work; and gaining further understanding of the scope and bounds of the group's professional interests. These steps are accomplished through two or three meetings with representatives of the group.

Once a group has been selected and its research interests identified, the appropriate literature must be selected and cataloged. For a meaningful catalog of limited size, literature selection is not a trivial matter. In reflecting the literature needs of the research groups, the journal and conference literature have been accorded nearly total emphasis. When a sufficient base of this prime literature has been established, other literature forms will be added to round out the catalog.

The research groups choose the set of journals of interest to them, and individual articles from this set, going back to 1 January 1967,

are selected. Article selection was initially performed by our librarians. An analysis of the librarian-selected, journal-article literature indicated that much material of peripheral interest to the group was being included. For example, in one test, the librarians selected 186 of 585 possible articles (31.8%), whereas a professor and a doctoral candidate in the research area selected 49 (8.4%) and 66 (11.3%) articles, respectively, from the same group of 585. Of those articles selected by the professor and doctoral candidate, the librarians had selected 73.5% and 63.6%, respectively, while the doctoral candidate selected only 46.9% of the articles selected by the professor. Apparently, the librarian's selection of a large proportion of relevant articles is achieved through the scatter effect of overselection. Accordingly, in order not to overextend the data base, a change was made in the article selection procedure. This change requires the cooperation and participation of two members of the research group. Copies of journal-issue tables of contents are routed to them; in return for this current awareness service, they indicate, independently, those articles they consider important to their area. All such articles, plus any obvious oversights caught by a librarian, are accepted for the data base.

A journal article selection card, shown in Fig. 9, is used for internal record keeping. On this card, an alphabetic code indicates which research groups have scanned the contents of a journal issue. For each article that has been selected, the initial page number of the article and the record number assigned to the article are entered on the card.

The text access experiments of Project Intrex utilize the same set of documents that are indexed in the augmented catalog. Full text of these documents is stored on microfiche. The fiche is prepared from microfilm. As a byproduct of the microfilming, full size electrostatic prints of the document are made and cataloging is done from the print rather than from the original document. This is important to our particular operation because the originals are not unduly detained from use by regular library patrons who may also be our potential system users. It also means that we can mark up the copies to suit our purposes. Microfilming is initiated through the 3-part, no-carbon-required form illustrated in Fig. 10. The last two parts are retained

Coden PHRVA		Physical Review										Vol. 161		Card 1	
Rec. No. 2786		Journal										Year 1967			
A B C D E F		A B C D E F		A B C D E F		A B C D E F		A B C D E F		A B C D E F		A B C D E F	
1		2		3											
090567		091067		091567											
117	1620	213	3636	569	3394										
189	1621	253	3637	637	3629										
194	1622	272	3638	736	3630										
202	1623	279	3639	756	3631										
102	1624	282	3640	815	3632										
133	1625	293	3641	852	3633										
172	1626	295	3642	877	3634										
179	1627	322	3643	903	3635										
NONE	C	350	3644												
		367	3645												
		434	3646												
		478	3647												
		493	3648												
		497	3649												
		506	3650												
		386	7100												
		398	7101												
		423	7102												

Fig. 9 Journal Article Selection Card, recording for each journal issue, the first page number of the selected articles and their assigned record numbers.

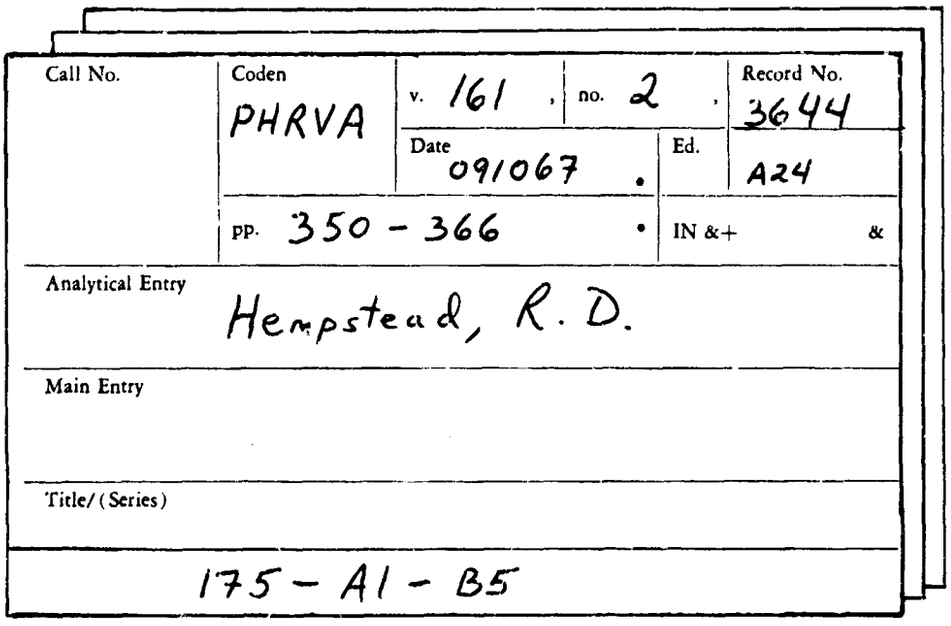


Fig. 10 Three-part no-carbon form used for microfilm initiation, for control files by record number and by analytical or main entry, and for transmitting data for fields 1, 2, 5, and 47.

for internal files ordered by record number and by analytic or main entry. A messenger uses the first part, or microfilm initiation copy, to locate the document, slip it, check it out of the library and deliver it to the Microreproduction Laboratory. Microfilming to the tolerances required by text access plus quality control checks take three days, at the end of which the prints of the microfilmed material are picked up and logged-in. The prints include a copy of the microfilm initiation slip; the original is kept by Microreproduction. The microfiche access number, assigned at the time of filming and recorded on the slip, appears on this copy. This number, when machine stored, will couple the text access experiments to the augmented catalog experiments. The slip print copy stays with the document print as it travels through the cataloging operation. The slip contains full formatted information for fields 1, 2, 5, and 47 which a typist later machine encodes.

B. RECORD CREATION

A cataloger takes one of the incoming document print copies and attaches to it the worksheet illustrated in Fig. 11. The left half of this sheet serves as a checklist of data fields which might be expected to appear for a given bibliographical form. The right half of the sheet is for input control information; each successive person processing the document indicates the task number for what they did, the date, and the time to accomplish the task.

In the cataloging operation, the intellectual task of indexing precedes the addition of descriptive data to a record. In order not to interrupt discussion of the overall work flow, details of the indexing procedure, per se, are deferred to the next section. The assigned weighted index terms are recorded on plain lined paper appended to the document print. Data for three other subject content information fields are supplied next: these are author's purpose, level of approach, and special features. The librarian, after checking a corporate name authority file and establishing any new corporate names associated with the document, then records data for fields 22 and 23. The authority file, currently on cards, ensures consistency in recording corporate names. We anticipate procedures for machine-storage of this file and the referencing of established corporate names

Field	No.	Art/Chap	Jnl/Srl Run	Monogr/Rpt	Thesis	Conf. Proc.
Record Number	1	X				
Document Selection	2	X				
Input Control	3	X				
On-line Date	4					
Access Number	5	X	X	X	X	X
L.C. Card Number	10		X	X	X	X
Library Loc. No. Copies	11		X	X	X	X
Serial Holdings	12		X			
Main Entry Pointer	20	X	X	X	X	X
Personal Names	21	X	X	X	X	X
Personal Name Affiliations	22	X	X	X	X	X
Corporate Names	23	X	X	X	X	X
Title	24	X	X	X	X	X
CODEN Title	25		X	X	X	X
Edition Statement	26		X	X	X	X
Publisher	27		X	X	X	X
Place of Publication	28		X	X	X	X
Dates of Publication	29		X	X	X	X
Medium	30	X	X	X	X	X
Format	31	X	X	X	X	X
Pagination	32		X	X	X	X
Illustrations	33	X	X	X	X	X
Dimensions	34		X	X	X	X
Serial Frequency	35		X	X	X	X
Language of Document	36	X	X	X	X	X
Lang. Accomp. Abstract	37	X	X	X	X	X
Series Statement	38		X	X	X	X
Report or Patent Nos.	39		X	X	X	X
Contract Statement	40		X	X	X	X
Supplement Referral	41		X	X	X	X
Errata	42		X	X	X	X
Thesis	43	X	X	X	X	X
Variants	44		X	X	X	X
Variant Title	45		X	X	X	X
Article Receipt Date	46	X	X	X	X	X
Analytic Citation	47	X	X	X	X	X
Abstract Services	48		X	X	X	X
Cost - A/C Source	49		X	X	X	X
Comm. Cost/Avail.	50		X	X	X	X
Author's Purpose	65	X	X	X	X	X
Level of Approach	66	X	X	X	X	X
Table of Contents	67	X	X	X	X	X
Special Features	68	X	X	X	X	X
Bibliography	69	X	X	X	X	X
Excerpts	70	X	X	X	X	X
Abstracts	71	X	X	X	X	X
Reviews	72	X	X	X	X	X
Subject Indexing	73	X	X	X	X	X
Article Ref. Citations	80	X	X	X	X	X
User Comments	85	X	X	X	X	X

1 RECORD 3644 -1

3

Worker/ Task,	Date St - Fin	Time St - Fin
		-
10/2	012968	1 30 - 2 25
		-
1/7	012968	2:40 - 2:56
		-
10/1	021468	140 - 150
		-
5/4	040168	11 35 - 11 50
		1 20 - 1 38
		-
		-
		-
		-
		-
		-
		-
		-
		-
		-
		-
		-
		-
		-
		-
		-
		-
		-
		-
		-
		-
		-
		-
		-
		-
		-
		-
		-
		-
		-
		-

/2 Cat S
/1 Cat D
/7 Review
/4 Key

Fig. 11 Cataloger's Cover Worksheet. Left side contains a checklist of data elements and the right side is for input processing control information.

by a number associated with the name. The librarian also supplies the data for fields 31, 39, 40, 42, 43, and 45 whenever these fields are applicable.

Another concern of the librarian at this time is the tagging of information on the document print for the typist's attention. The tagging includes: adding relators to personal names; decisions on inclusion of excerpts (appropriate text is circled) or tables of contents; and pointing to hidden data, as might be the case for the date an article was received for publication. The tagging procedure is in effect for journal and serial articles and conference papers; for other document types, the librarian currently supplies all the descriptive cataloging data.

The document and its record are then reviewed by another librarian. Additions or corrections to the record may be made. This review is primarily a check on the indexing operation and it is discussed more fully in the next section. After review, the documents and their records proceed to the typists.

The typists are responsible for gathering and formatting other descriptive cataloging data from the tagged copies of articles and conference papers; they also gather from the worksheet and format the input control information for all document records. This information, together with the librarian-generated information, is typed and punched.

Friden 2303 Flexowriters are used to produce machine-readable punched paper tapes of the catalog records simultaneously with typed hard copies. The paper tape mode was selected because it accommodates a large character set, including upper and lower case, as well as ease in handling lengthy records containing many variable length data fields.

As a practical matter, the last field entered on a record is field 3, input control information, so that the typist may enter the time taken to machine encode the record. Typing errors are corrected on the paper tape only when the typist immediately catches the error. Other errors that the typist detects in the record are marked on the hard copy for later correction.

At the typing stage, ten records are batched to form one file of paper tape input. The first record in a file is preceded on the tape by

the file number assigned by the typist. The last record in the paper tape file is followed by a punch code for "stop". Subsequent input, error identification, and error correction operations are on the basis of a file. All hard copy documents and records are temporarily stored by file; the files of tape are logged out and turned over to software personnel for computer input.

The paper tape file is read into an IBM 7094 computer through a satellite PDP-7 computer, and Flexowriter codes are converted into ASCII codes. The file, now called a working file, is stored on disk, and printed on a 1403 line printer, equipped with an extended character-set print chain. The printout is returned to the input group and matched with the hard copy file.

C. ERROR IDENTIFICATION AND CORRECTION

Proofreading of the first computer printout of a file against the previous hard copy is performed by a librarian. This is also the first time that the descriptive cataloging performed by the typists is checked. Consistency among certain fields is also sought; for example, if data appears in field 43 for a thesis, associated personal and corporate names with appropriate relators should also appear in fields 21 and 23, respectively. Errors are marked for correction on the printout and a separate tabulation is made of the number of errors in each of four classes (cataloging, policy, typing, mechanical).

Correction of errors in the computer-stored working file is done by a typist at an IBM 2741 console using an on-line context editing program. On-line context editing allows a dialog between typist and computer in identifying and verifying changes in the data by referencing the changes to be made by their context, or surroundings, in the file. A sample extended dialog is shown below.

/within the framework/ (1)

within the framework of the simple convering collision-time model. (2)

v/convering/conserving/ (3)

within the framework of the simple convering collision-time model. (4)

s (5)

In the sample dialog, the typist specifies the line to be corrected by typing a suitable unique portion of that line (see line 1 above). After the computer finds a line having that specification, it responds by typing the entire line (see line 2). The typist then determines that it is indeed the desired line. If it is, the typist specifies the characters to be changed and the manner of the change. On line 3 of the example, "convering" is to be changed to the correctly spelled word "conserving". The "v" appearing at the beginning of the sequence is a command indicating that a verification is to be made before the character substitution. The computer responds by retyping the entire line (see line 4) and emphasizing, by typing in red, those characters which the computer understands are to be corrected; these are the characters underlined twice on line 4. If this is the correct substitution, the typist next enters the command "s" for substitute (see line 5) and the change is made. Use of the verify command before the substitute command guards against a data change on the basis of an ambiguous context specification. In addition to making substitutions, the typist can also delete lines or insert new lines.

During the on-line editing, additional errors in the file not previously caught in proofreading are also corrected, and a separate tabulation is made of their number and class. A new printout of the edited working file is made. This printout is proofread by a typist against the previous printout. Any errors are corrected on-line. When no further errors are detected in a file printout, the on-line editing program is used to certify that the records in the file are ready for further processing into the computer-stored data base, that is, their status is changed. Computer processing of individual catalog records now replaces the computer processing of files of batched catalog records.

D. COMPUTER FILE ORGANIZATION

The corrected disk-stored catalog records are restructured and to each record a header table of field locations is added which replaces the field tags. Additionally, information is extracted from the records to create two disk-stored inverted files to the set of catalog records, that is, a personal name index, and a combined subject-title index, each with references to catalog record numbers. Each reference

contains, in addition to a record number, codes for certain attributes of the associated document. For example, in the subject term inverted file, reference attributes might indicate whether or not the document is a journal article written for a professional and containing results of original research. Attribute codes are automatically assigned from the data extracted from the record.

Both of the inverted files and the complete catalog record files will each have a directory serving to localize the position of file entries on the disk. The computer file organization is shown in Fig. 12. The catalog record file directory will reside on disk, but the inverted file directories will reside in core memory. This computer file organization permits three levels of searches: searching on inverted file keys (personal names, or subject/title terms); searching on specific attributes of documents as given in the references listed under inverted file keys; and searching through the fields of the catalog records themselves.

Concurrent with the creation of the inverted files and the associated directories, two special files are also created. One is a bookkeeping file containing control information data extracted from field 3 and intended for use in statistical compilations on document processing rates. The second special file is a rejected records file containing records which computer programs recognize as containing potential errors; human tracing of these "errors" is required to remove the objection to the record raised by the computer program.

E. INDEXING

The indexing process requires the greatest amount of the professional effort expended in cataloging. Our current procedures call for the use of terms based upon the text of a document. In general, terms are combinations of phrases. A term from a document may require further intensification to provide fairly complete context among its individual components. Each term is structured to provide sufficient contextual expression such that the term may stand by itself. Further, each term is weighted to reflect that proportion of a document devoted to discussing the represented concept. Illustrations of the set of subject terms for two records may be seen in Figs. 2 and 3. There is

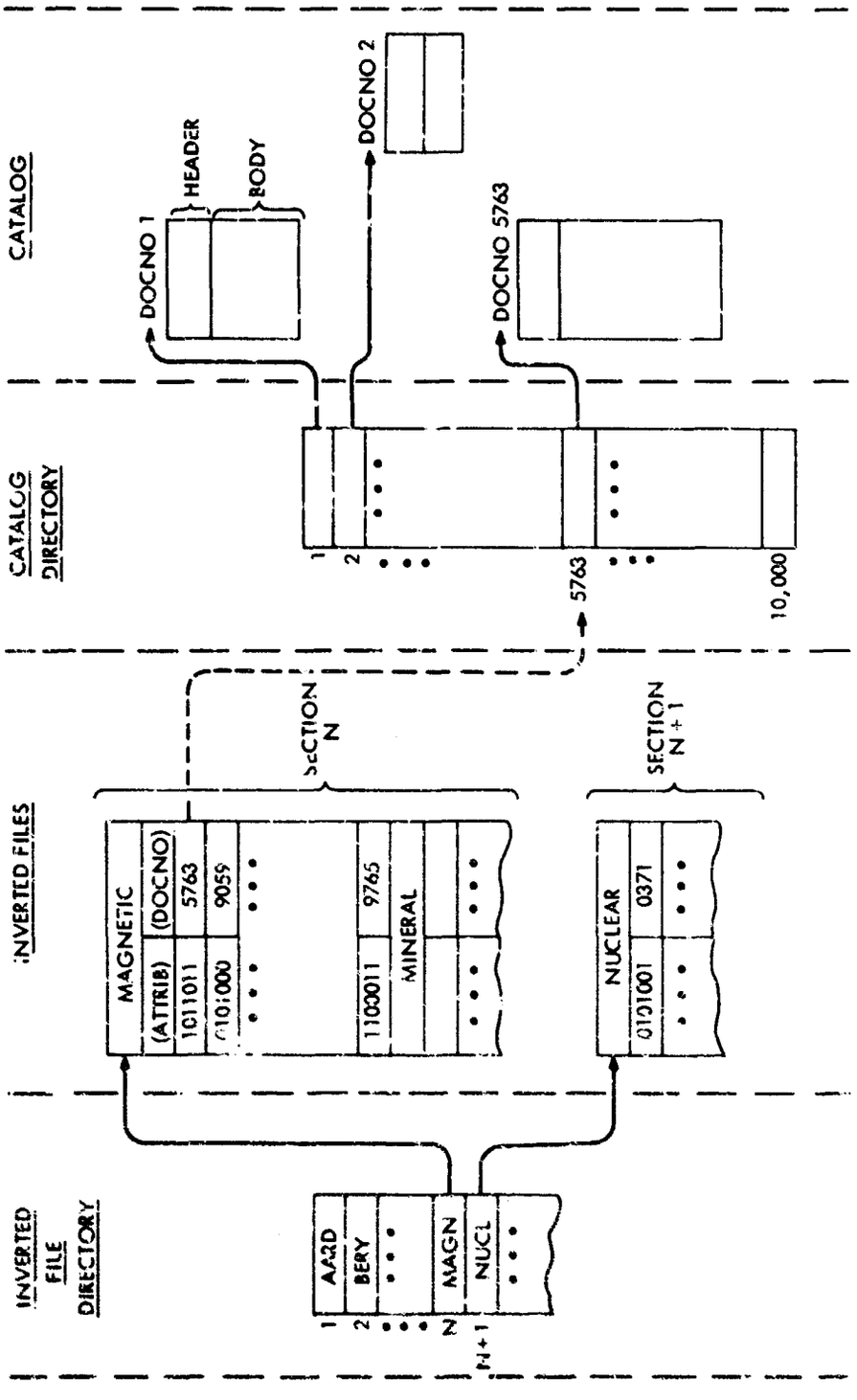


Fig. 12 Computer File Organization (simplified diagram)

no limit on the number of terms assigned to a document and no authority list of terms is used.

There are five steps in the indexing process:

1. Scanning a document to get an idea of its coverage.
2. Identifying those concepts in the document which are to be indexed.
3. Formulating an appropriately structured subject term to represent the concept; words and phrases not appearing in a document may be used.
4. Weighting the subject term.
5. Recording the final decision.

These steps are often performed simultaneously.

Good initial sources of subject terms are a document's title, abstract, section headings, introduction, conclusions, and illustrations and their captions. These sources can be insufficient, in which case text must be examined in sufficient detail beyond a general scan. Text provides additional sources of whole subject terms or additional words for term intensification. Terms receive weights of one when they represent most of the content of an entire document, weights of two when they represent most of the content of an entire section of a document, and weights of three when they represent the content of only small portions of a document. Materials, tools, techniques, and applications, which deserve to be indexed but which do not otherwise appear in any other term, are accorded weights of four. Terms of a more generic nature to the document may be added and these terms receive a zero weight.

In the review operation, another librarian provides a quality control on indexing by checking the work of the original cataloger. Points looked for include: relative completeness of the set of terms in covering the total essential concepts expressed in a document; completeness of each individual subject term in expressing a concept; redundant terms; trivial terms; awkward expression; readability; accuracy; appropriateness of a weight. The reviewer may make whatever deletions, changes, or additions to the indexing are deemed necessary. Only the very major changes are discussed for feedback

purposes with the original indexer. In all cases, the reviewer's decision is the final one. Because all the librarians serve both as indexers and reviewers, an important indexing information exchange, only partly verbal, occurs during the review. Indexing is a technique learned through continuous experience. The review serves as a common educating device by increasing the reviewer's awareness of the total indexing forest; techniques and pointers gained by a reviewer from the work of others can be applied to his own future indexing of the trees.

The indexing technique is not a rapid one but, as we shall see shortly, it allows flexibility in designing experimental information retrieval systems. In order to increase the rate of cataloging without sacrificing the basis of the indexing procedure, an indexing time limit and a student indexing program were instituted.

The time limit, based upon the early experiences of the indexers, is correlated to the number of pages in the document being indexed; 20 minutes are allowed for indexing documents containing up to three pages, 30 minutes for documents containg four or five pages, and on up to a maximum of 75 minutes for a document of 20 or more pages. The institution of such time limits has the added advantage of creating an immediate fixed goal for the indexer to meet. A goal of x documents indexed per time period does not account for document length and that is potentially detrimental to the indexing.

Students, predominantly undergraduates in science and engineering, are being employed to index documents. Prior to assuming indexing duties, these students attend three training sessions and receive practice indexing assignments for homework following the first two sessions. The training program initially orients the students to the scope of Project Intrex and the nature of the augmented catalog experiments before proceeding to cover the fundamentals of their specific job of indexing. During the training sessions and in discussions of their assignments, the students are presented with a spectrum of acceptable indexing, and they are encouraged to develop their own consistent, yet acceptable, approach to indexing. At their last meeting, each student is assigned to a librarian who will serve as his reviewer. No more than two students are assigned to a librarian.

Further training and guidance of the student in indexing is provided on the job by the reviewer. Although encouraged to do their work at the laboratory, the students may work at home; each student is expected to work about ten hours a week.

The cataloging for which each student is responsible is the indexing of a document (field 73), the author's purpose (field 65), and the level of approach (field 66). The student's reviewer critiques the indexing more thoroughly than would be done for another librarian's indexing. Corrections, additions, or deletions to the indexing are analyzed and these are discussed with the student, generally once a week. The reviewer also adds other necessary fields (for example, corporate names) to the student-initiated record before it enters the normal work-flow pattern at the typing stage.

Our first experiences in using students as indexers have been varied, but we consider the overall program to be successful. The rapport between student and reviewer has been generally very good. Each provides the other with a better insight and perspective of the nature of the indexing process. Most of the students have produced acceptable indexing. Some students find that they cannot devote the expected number of hours to our project. Because indexing is a technique learned from experience, the quality of their work suffers and these students are eventually dropped from the program.

F. SUBJECT TERM RETRIEVAL

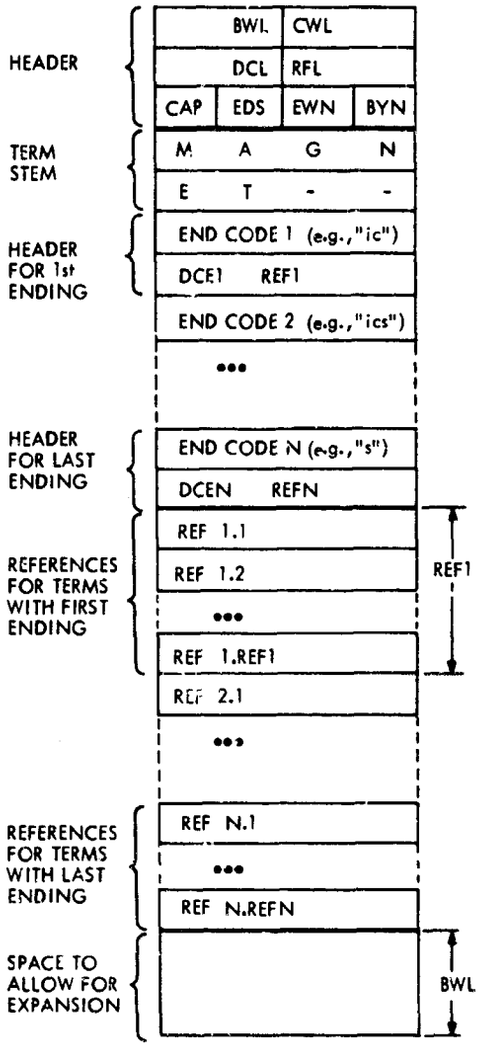
The indexing operation is basically one of wide freedom in choice of terms and their component words; in practice, emphasis is given to words used by an author. A user initially approaching the catalog will have his own set of terms which may be different from the set used by an author and indexer in describing a concept. At Project Intrex, we do not wish to constrain either an indexer or a user into using any relatively fixed set of terms as represented by subject authority lists and some thesauri. How then do we try to reasonably bridge the vocabulary gap between indexer and user? The answer lies in the software programs which manipulate both sides of the index terminology-user terminology interface.

It is unlikely that full matching will occur between user terms and index terms when more than two or three words constitute the term.

Consequently, searching strategies are based on partial term matching as well as full term matching. In order to efficiently effect this strategy, the subject term inverted file will contain as entry keys both the full subject term, and phrase decomposed single words from each term. All the words appearing in a single subject term are linked by virtue of their appearance within the one term. Because the subject terms are structured to be complete and readable phrase expressions, the word order and syntax of a term play a role-like effect without the overly rigorous constraints of role indicators. Numbers indicating the order of a word in a term, and the order of that term in the term set from a record, are entered as attributes for each reference given under that word or key term in the inverted file. Through this mechanism, searches can be made for individual words (derived from strings of words) while also specifying particular context or association with other words at a level smaller (or larger, if the full set of terms is considered) than that of one complete subject term.

A word may have one of a multiple number of endings depending upon syntax and context of its use in a subject term. For example, the stem "magnet" may have the endings "s," "ic," "ism," etc. Consequently, words stored as keys in the inverted file are reduced to their stem, and for each stem the word endings serve as subkeys. Searching can be initiated for full words or for stemmed words. Syntax can be considered or ignored. The organization of the subject term inverted file is shown in Fig. 13.

Stemming and phrase decomposition are applied not only to the index terms but also to the subject terms in a user query. If partial matches through these two techniques are the order, when will the complete subject term expressions as furnished in indexing be used? As we have seen, the full terms provide certain attributes to the phrase decomposed words. Additional aids that utilize the full term relate to the display function capability of a cathode ray tube console at the man-machine interface. For example, an initial partial match might yield an excessive number of potential references to documents. One technique of further narrowing the search is the user's response in a man-machine dialog which asked him to select the more appropriate contexts from a displayed list of full subject terms initially selected by partial matching. In another example, a user who wisher to



BWL: Number of blanks at end of list (for expansion)

CWL: Total number of computer words on list

RFL: Number of references

DCL: Number of distinct documents among references

BYN: Number of bytes in term stem (here 6)

EWN: Number of English words in term (here 1)

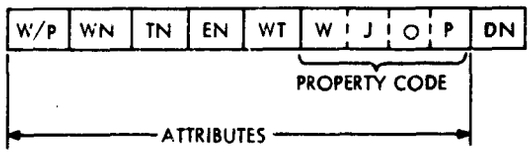
EDS: Number of different endings (here N)

CAP: Bit to indicate variable capitalization

REF1: Number of references for the work "magnetic"

DCE1: Number of distinct documents referring to "magnetic"

REFERENCE-WORD FORMAT



DN: Document number

W/P: Is term a single word (W) or full phrase (P)?

WN: Word number within phrase (for W/P = W)

TN: The term number of this term for given document.

EN: Word ending number for this reference (from 1 to N)

WT: The subject/title weight (level).

W: Is document whole work?

J: Is document journal article?

O: Does document reflect original work?

P: Is document written for professional?

Fig. 13 Format for the Inverted File Subject-Term List

further discriminate among a set of references to documents could call for display of any data field in a catalog record. In this manner, display of the set of subject terms provides the user with a more complete picture of the document's contents.

What about synonyms? This is a problem and one proposed solution is a thesaurus. Won't this constrain the indexer and user in their terminology? No, because the thesaurus will not also be an authority list; rather, it will be a machine-stored guide and cross reference switching device. A user will be able to see displays of term relationships. He will have freedom to use thesaurus suggestions as he sees fit but he will also be able to specify automatic computer searching on terms synonymous with his term. Automatic searching from this thesaurus would be akin to searching on terms indicated in contemporary thesauri by "use," "used for," and "narrower term." This thesaurus will be part of a later, more sophisticated phase in the augmented catalog development. Its construction may be based on index terms and user terms in the earlier phases of the catalog's development and on machine-aided techniques used in studies of automatic indexing and the co-occurrences of terms.

IV. EVALUATIONS

Several initial studies were conducted to evaluate the catalog data input. It has been our aim in these studies, which are briefly described below, to delineate characteristics of the catalog records and of the data input processes, to point out particular operations used for improvement, and to identify potential areas for further detailed study.

Our original processing of catalog records included the local generation of reference citation data for field 80. A typist standardized the format of the references contained in journal articles and punched a paper tape of this data; at a later stage the citation data tape was merged with the main catalog data tape to produce a single paper tape. An evaluation of the processing of our first 50 records (February-March, 1967) revealed that the time required for the citation formatting and keying processes alone (89.0 man-minutes per record) was comparable to that required in the generation and keying processes of the main catalog record data (82.7 man-minutes per record). This data is summarized in Fig. 14. If local generation of machine-readable citation data were to continue without a backlog, personnel equivalent in number to those working on the generation of the main record would have had to be hired. We concluded that a more logical solution would be to suspend local generation of citation data and to seek machine-readable citation data from outside sources. Discussions have been held with the Institute for Scientific Information in Philadelphia concerning the purchase of citation data.

The initial start-up set of operations (see Fig. 14) included, for example, complete cataloging by the librarians (now the typists handle most of the descriptive cataloging), keying, proofreading, and editing of individual records (now all handling in these operations is on the basis of a file of 10 records), a proofreading-editing loop using successive generations of paper tape (now this loop occurs <u>after</u> initial computer input and the editing is an on-line operation). These major changes in our operation were implemented as we gained better understanding of the nature of each job and as increased computing facilities became available to us. The current workflow, illustrated in Fig. 8, is considerably different from the initial start-up workflow. The average

AVERAGE TIME OF INITIAL
INPUT WORKFLOW OPERATIONS

Process	Time per Record (man-minutes)	
Descriptive Cataloging	10.6	
Subject Cataloging	23.3	
Review	6.4	
Keying	33.9	
Proofreading	8.5	
Subtotal		82.7
Citation Formatting	48.8	
Review	4.2	
Keying	31.8	
Proofreading	4.2	
Subtotal		89.0
Merge and Editing (of above two tapes)	27.6	
Proofreading	4.2	
Editing	8.5	
Subtotal		40.3
TOTAL		212.0

Fig. 14 Average Times of Initial Input Workflow Operations for the First Fifty Records Processed (February - March, 1967)

processing time for each current operation (January-April, 1968), and related data, is given in Fig. 15.

Initially, the average time taken to catalog, review, key, and first proofread a record was 82.7 man-minutes; the comparable current operations take an average 68.1 man-minutes per record. The difference is primarily due to the cut by half in the time to key a record. The typists' increased proficiency is all the more significant because they now have the additional responsibility for formatting most descriptive cataloging elements. In current processing, 80 percent of the files cycle through two edit loops (once for editing and once for certification); 20 percent of the files have errors to be corrected in a second on-line editing step and so these files are cycled through three edit loops. About half of the 11.2 average number of errors per file detected during a first proofreading are typographical errors; cataloging errors comprise the remainder. As the number of errors per file decreases, the ratio of total computer-time consumed to console-time consumed in on-line editing increases. The ratio is 1/36 for the first on-line edit of a file and it is 1/6 for on-line certification of an error-free file.

The data presented in Fig. 15 is on a per-record or per-file basis. We can gain further perspective in this analysis by considering some average length parameters of a logical catalog record. There are an average 2461 keystrokes per record, exclusive of carriage returns, tabs, and case shifts, or 410 six-character English words per record. There are 16.4 index terms per record, averaging 7.7 English words per index term in length. At present, the data base consists almost entirely of journal articles or conference papers, and in the survey sample these documents averaged 5.3 pages in length. The set of index terms comprises about 1/3 the length of a catalog record, an abstract or excerpt comprises about 2/5 of the record length, and all other applicable data elements account for the remaining catalog record length.

We have made an initial study of the economic feasibility of on-line teletypewriter terminal keying with respect to off-line paper tape keying of the cataloging data. Six files of catalog records were keyed on-line by a typist working at an IBM 2741 terminal connected to the 7094 computer. The on-line input was cost compared to that for six files input by the paper tape medium. The analysis of each input mode included

CURRENT AVERAGE PROCESSING TIMES
and
ASSOCIATED AVERAGE CHARACTERISTICS

Process	Average Time
	(man-minutes/record)
Indexing	28.0
Descriptive Cataloging (by Librarian)	4.7
Review	10.3
Keying plus Descriptive Cataloging (by Typist)	16.9
	(man-minutes/file)
Proofreading (by Librarian)	82.0
On-Line Editing	20.7
Second Proofreading (by Typist)	11.5
Second On-Line Editing (20% of the files)	4.1
Third Proofreading (by Typist) (20% of the files)	1.7
On-Line Certification	1.0

Associated Characteristics	Average
Computer Time/Console Time Ratio	
First On-Line Edit	0.028 (1/36)
Second On-Line Edit	0.074 (1/14)
On-Line Certification	0.17 (1/6)
Errors Caught During	(per file)
First Proofreading	11.2
First Editing	0.33
Second Proofreading	2.0
	(per record)
Keystrokes	2461
English Words	410
Index Terms	16.4
English Words/Index Term	7.7
	(pages)
Document Length	5.3

Fig. 15 Average Processing Times of Current Workflow Operations and Associated Average Characteristics (January - April, 1968). One file contains ten records.

all of the computer operations through the generation of a first printout. The data is summarized in Fig. 16; it is valid only for the current (March 1968) M.I.T. compatible time sharing system environment. The unit cost figures for typist salary (including overhead), and keying machine rental costs, are pro-rated over 20 working days per month, seven hours per day. Materials associated with on-line input are negligible and there is no direct charge to us for printout. The analysis shows that the cost of on-line input of data to the augmented catalog in the current CTSS environment is $41.07 per file whereas the cost of paper tape input is $23.46 per file. The cost differential is due almost entirely to the computer time and cost required to process data input from the on-line terminal (205.4 seconds versus 6.4 seconds). All other operations in the two input modes are nearly equivalent in time and cost. As the hardware associated with the Intrex system becomes available, and as the computing system environment changes, on-line input will be restudied; particular attention will be given to the effects of using a local buffer controller which is, in turn, connected to the larger computer system.

Another initial study[8] was made to determine the effects of indexer experience over time on the subject indexing time per record. The data is scattered but there is evidence of an average learning period of three months duration for librarians and of two months duration for students. At the end of this learning period, indexing time per record (averaged over all indexers) levels off at six to eight minutes per page from an initial high of ten to twelve minutes per page. In another phase of this study, the average indexing time per page was found to be influenced by document-related parameters. There were observable differentials of the per-page indexing time among documents grouped by format. For example, the ease of indexing articles decreases in this order: conference paper, regular journal article, letter-type journal article.

COST ANALYSIS OF ON-LINE INPUT AND PAPER TAPE INPUT

Operation (Unit Cost)	On-Line Input		Paper Tape Input	
	Time per File	Cost per File	Time per File	Cost per File
Keying Operator	(minutes)		(minutes)	
Preparation	27.0		30.0	
Log In	2.4			
Input	174.3		164.2	
Total Keying ($0.095/min.)	203.7	$19.35	194.2	$18.45
Computer	(seconds)		(seconds)	
Log In Subtotal	8.4			
Processing	7.1			
Swap	1.3			
Input Subtotal	205.4		6.4	
Processing	54.1		3.4	
Swap	151.3		3.0	
Magnetic Tape Generation, Printout and Miscellaneous Operations Subtotal	6.5		8.0	
Processing	5.6		6.8	
Swap	0.9		1.2	
Total Computer ($0.083/sec.)	220.3	$18.28	14.4	$1.20
Computer Operator ($0.055/min.)	(minutes) 3.2	$0.18	(minutes) 7.0	$0.39
Keying Machine Rental	(minutes)		(minutes)	
IBM 2741 and associated equipment ($0.016/min.)	203.7	$3.26		
Flexowriter 2303 ($0.014/min.)			194.2	$2.72
Materials				$0.70
TOTAL COST		$41.07		$23.46

Fig. 16 Cost Analysis of On-Line Input and Paper Tape Input of Catalog Data to a 7094 Computer Operating in the MIT CTSS Environment (March 1968). One file contains ten records.

SUMMARY

A flexible analytically structured catalog record format was designed to aid in meeting the objectives of the display-oriented Project Intrex augmented catalog experiments. The catalog data elements and their encoding for machine readability have been discussed. The selection of documents from the literature of materials science and engineering for the Intrex data base, the generation of catalog records of those documents, and the initial processing of those records for computer-storage were covered. Initial studies were made to evaluate several aspects of the data and the data input processing. One study has shown that to input data at an on-line terminal in our current MIT CTSS operating environment is twice as expensive as our normal off-line data input using punched paper tape. Attention was given to the creation from each document of a set of complete index term phrases and to the problems of matching these unconstrained terms with similarly unconstrained subject request phrases. Computer programs for phrase decomposition and word stemming, and interactive man-machine dialog, will help solve the problems of subject retrieval. The main development phase of the experimental time-shared augmented catalog is nearing completion, after which, active experimentation on the Intrex system with user groups will begin.

REFERENCES

1. Overhage, Carl F. J., and R. Joyce Harmon, eds. Intrex, report of a planning conference on information transfer experiments. Cambridge, Mass., M.I.T. Press, 1965.

2. Details of the evolving research program may be found in: Massachusetts Institute of Technology. Project Intrex. Semi-annual activity reports. Cambridge, Mass. (PR-1, 15 March 1966, and continuing)

3. Benenfeld, Alan R., Elizabeth J. Gurley, and Jane E. Rust. Cataloging manual. Cambridge, Mass., Electronic Systems Laboratory, Massachusetts Institute of Technology, February 1967. (ESL-TM-303).

4. Avram, Henriette D., Ruth S. Freitag, and Kay D. Guiles, A Proposed Format for a Standardized Machine-Readable Catalog Record; preliminary draft. Washington, D. C., Library of Congress, Office of the Information Systems Specialist, June 1965. (ISS Planning Memorandum 3).

5. ----- -----. Supplement 1. October 20, 1965.

6. Atomic Energy Commission. Division of Technical Information Extension. Descriptive Cataloging Guide. 1st revision. Oak Ridge, Tenn., January 1966. (TID 4577 (Rev. 1)).

7. Curran, Ann T., and Henriette D. Avram. The Identification of data elements in bibliographic records. May 1967. (Final report of the Special Project on Data Elements for the Subcommittee on Machine Input Records (SC-2) of the Sectional Committee on Library Work and Documentation (Z-39) of the United States of America Standards Institute).

8. Lufkin, Richard C., Determination and Analysis of Some Parameters Affecting the Subject Indexing Process. Bachelor Thesis. Massachusetts Institute of Technology, June 1968.