MULTIPURPOSE CATALOGING AND INDEXING SYSTEM (CAIN) AT THE NATIONAL AGRICULTURAL LIBRARY.

Vern J. VAN DYKE: Chief, Computer Applications, National Agricultural Library, and Nancy L. AYER: Computer Systems Analyst, National Agricultural Library, Beltsville, Maryland.

A description of the Cataloging and Indexing System (CAIN) which the National Agricultural Library has been using since January 1970 to build a broad data base of agricultural and associated sciences information. With a single keyboarding, bibliographic data is inputed, edited, manipulated, and merged into a permanent base which is used to produce many types of printed or print-ready end-products. Presently consisting of five subsystems, CAIN utilizes the concept of controlled authority files to facilitate both information input and its retrieval. The system was designed to provide maximum computer services with the minimum of effort by users.

INTRODUCTION

This article describes an interactive system in operation at the National Agricultural Library which with a single keyboarding of data provides all necessary catalog cards, book catalogs, bibliographies, and related internal reports, as well as a computer data base for information retrieval. Primarily in batch mode, the system can operate on an IBM 360 with 256K memory using OS, six magnetic tape drives, a card reader, and a line printer.

BACKGROUND

The National Agricultural Library (NAL) as one of the three national libraries is responsible for the collection and dissemination of agricultural information on a national and worldwide basis. In this pursuit publications are obtained through gifts, exchange agreements, and by purchase of items in many languages. Titles of those items in non-Roman alphabets are transliterated and all non-English titles are translated.

The volume of publications handled by NAL in 1969 was in the neigh-

borhood of 600,000, of which approximately 275,000 were added to the collection. This volume was sufficiently large to provide a serious problem to NAL's staff and thus computer assistance was clearly a logical and necessary arrangement.

In 1964 a computer group was formed in NAL; it became active in developing systems to prepare voluminous indexes for the *Bibliography of Agriculture,* the complete *Pesticides Documentation Bulletin,* and the categorical and alphabetical issues of the *Agricultural/Biological Vocabulary.* During 1969 these systems were consolidated and expanded so as to process all input data within one coordinated set of parameters. In January 1970 the new Cataloging and Indexing (CAIN) System was implemented.

SYSTEM DESIGN

CAIN is a complex and comprehensive computer system which has been engineered to handle up to five (5) simultaneous but separate users who share the same controlled authority files. The basic precept in development of computer applications at NAL is to make input and output simple and convenient for the users, with the computer assuming as much detail and data manipulation as is technically feasible. At NAL the current users providing input data are the New Book Section, Cataloging, Indexing, and Agricultural Economics. Operating in parallel, CAIN also services the herbicides data base of the Agricultural Research Service; the International Tree Disease data base of the Forest Service; and in 1971 will be installed in the Library of the Technion-Israel Institute of Technology in Haifa, Israel.

The master data record is variable in length with a fixed portion of 173 characters and up to fifty-seven additional segments of 65 characters each. The fixed portion includes basic data plus a directory of data contained in the variable portion. Data elements in CAIN are:

a. File code—delineates the various files.
b. Identification number—on cataloged items this embodies the accession number. All identification numbers include the year of accession, a parallel run code plus a unique control number.
c. Source code.
d. User codes—specific identification of up to five users.
e. English Indicator—language of text.
f. Translation code—availability of an English translation.
g. Language, if other than English.
h. Proprietary restrictor—identifies classified records.
i. Title tracing indicator—for catalog cards.
j. Main entry—designates main entry if not normal sequence.
k. Document type—whether journal article, monograph, serial, etc.
l. Filing location—if other than in the library stacks.
m. Categories—two. General area of coverage of subject matter.

n. New book description—if the title is not sufficiently explanatory.
o. Titles—three types: (1) vernacular or short, (2) alternate or holdings, and (3) translated title (English).
p. Personal authors—up to 10. Names plus identifying data.
q. Corporate authors—maximum of two.
r. Major personal author affiliation.
s. Abbreviated journal title if item is a journal article; imprint if monographs and serials.
t. Collation/Pagination.
u. Date—two: Search date, and date on publication if different.
v. Call number.
w. Subject terms—may be nested. Up to 45.
x. General Notes.
y. Special purpose numbers—patent, grant, analysis, contract, technical, or report.
z. Series statement.
aa. Abstract/Extract.
bb. Tracings not otherwise normally generated by the system.
cc. Nonvocabulary cross-references.

The total number of individual elements is limited only by the maximum record size.

The NAL-produced software is written in COBOL. The data base is maintained on tape which is nine-track, 800 bpi, blocked 2, in EBCDIC, with standard IBM 360 header and trailer labels. The total system presently consists of forty programs, some of which are multipass. In addition, throughput is sorted twenty-five times during the full computer run. These, of course, include the search and retrieval programs and sorts which are run only on request.

The ultimate system which NAL is working toward and for which the basic design is already substantially complete is an on-line full library document locator and control system which may be linked via dial-up service to an international and national science and technology information network. Each portion of CAIN is developed with the broader picture in mind. It was this factor which weighed heavily in selecting cathode ray tube (CRT) terminals for the proposed data gathering subsystem inasmuch as CRT's will be the predominant type of terminal in the future network.

For convenience in discussion, the system will be described by its subsystems: data gathering, edit and update, publication, search and controlled authorities.

DATA GATHERING SUBSYSTEM

From its inception the input to CAIN was in the form of punched cards, a method which has proved to be slow and error prone. In order to eliminate double keyboarding and excessive time lag, as well as to reduce the

error rates, it was decided to perform this input function in the library with trained library personnel. To accomplish this, NAL proposes to implement an "on-line" type of input subsystem using CRT's. Although this form of entry is not yet in use, the subsystem should operate substantially as follows.

The documents are to be marked by catalogers and indexers and passed to library technicians who will enter the data through CRT's into an on-line storage file. To do this, the technician will call from the hardware prestored formats as desired and fill in the data elements required. These formats use English terms and for the most part call for data rather than codes. In addition, data are to be entered in normal upper- and lowercase without diacritics, thus improving visual scanning for errors. An average of four formats will be needed to enter one item. By use of an algorithm, the system would store formatted records for each ID in such a manner as to permit recall singly or collectively.

The physical documents are then to be passed on to an editor who can recall any or all formatted records for review. With the document in hand, stored records will be reviewed and corrected if necessary. When acceptable, the records will then be transmitted to magnetic tape.

Variations on this procedure could include input direct to tape, storage to tape without recall to a CRT by an editor, cancellation of actions, and a direct purge of the entire storage file without loss of the controlling matrix.

The expertise of the library technicians inputting the data should insure far more accuracy than could be expected from multihandling and multikeyboarding. In addition the system has been designed to accomplish basic pre-CAIN editing of such factors as numeric or alphabetic characters in certain fields and overall lengths of the fields. Errors in these categories will be promptly identified by the computer by a blinking feature on the CRT screen.

Another major benefit of this direct approach is that documents can be processed through the system so as to reach the stacks twenty-four days faster than under the current keypunch method.

Magnetic tapes created by the data gathering system will be periodically converted from ASCII to EBCDIC and processed into the edit and update subsystem of CAIN. The present NAL time schedule for updating master CAIN files is weekly. This is not a requirement of the system but an administrative decision based on other deadlines.

The data gathering system as prescribed by NAL will be composed of sixteen CRT's, a large on-line storage file, and one nine-track 800 bpi magnetic tape drive. This configuration will be either a hard-wired "black-box" approach, or controlled by a dedicated mini-computer. The hardware prescribed for this subsystem is not included as a requirement of CAIN inasmuch as transactions can be entered on 80-column cards if desired.

An additional feature of this subsystem will be the generation of manage-

ment information feedback. This will encourage elimination of manual counts and provide accurate throughput volume statistics on a timely basis. Through this means the supervisor will be in a better position to evaluate workload, individual performance, and hardware utilization.

EDIT AND UPDATE SUBSYSTEM

The first step in the acceptance of transactions is a thorough validation of each data element. The computer is used to relieve librarians of the voluminous and time-consuming edit of many individual elements having predetermined limits. Thus, only a cursory review of the proof-listed records is necessary by a librarian before acceptance. The system cannot detect, of course, logical or typographical errors, but it can determine the absence of necessary information, codes in invalid ranges, and the incorrect placement of data.

Elements for which the system supplies authority files are not only verified against the file but also additional transactions are generated from the authority file to assure uniformity in output. This also eliminates the necessity for librarians having to enter those elements which have a direct predictable relationship to another element.

Further validations are performed at the point of building new records or updating records already in the master file. The two "master" files are (1) the temporary set of unselected records and (2) the permanent set of those records which have been approved and selected for publication in some form. Data elements specified as required within each record are reviewed. If one or more is missing, the system refuses to approve this record, and a notice is produced concerning this reversal of human input.

Fields can be deleted, in whole or in part, replaced or added.

Three types of output from this subsystem are:

- New updated master files. Those which have been added or altered during this update run are proof-listed for cursory review by a team of professional librarians. Corrections and/or approvals are submitted in a subsequent update run.
- Activity notices. Every action whether submitted by the user or system-generated which has been accepted for processing is reported.
- Error notices. All error and warning messages from this subsystem are compiled into one listing. This includes errors on individual elements, system-discovered errors of omission, and warnings of computer overriding of submitted actions.

Through the use of control cards various handling options are possible. One of these is proof-listing of a specific range or ranges of masters by identification numbers or dates.

Subject headings are assigned by professional librarians for monographs and new serial titles. For journal articles, however, the system analyzes the title of the article and creates subject index terms, using single words,

combinations of two words not separated by stop words, and singular and plural variations. The generated terms are then processed against the controlled authority file. Those accepted as valid are inserted in the record for searching purposes.

PUBLICATION AND DISTRIBUTION SUBSYSTEM

Each data element of a bibliographic item is captured only once and at the earliest possible time in the receipt process. Master records which have successfully passed the edit and update phase become candidates for various types of publications and other user services.

Six major modes of publication products are produced by CAIN, at various times and in a variety of both formats and media.

Preliminary to the production of formal output there is a screening for records designated as fully acceptable by the edit and update subsystem. As mentioned above, any record may be identified as being applicable to any combination of from one to five users. By a method of control cards the system is informed as to which users are scheduled for publication/distribution, and the maximum quantity to be selected in each case.

This subsystem reviews each record to ascertain its appropriateness for selection. Records meeting the criteria are siphoned off for individual handling. No record is dropped from the temporary file until it has been selected by all applicable users.

A New Book Shelf listing may be printed on photocopy paper on request. On preparation, it is ready to be matted, photographed, printed, and distributed throughout the Department of Agriculture. Only enough new book entries are selected by the computer at one time as will fit on three sheets of a four-page publication.

Approved cataloged records are selected weekly. Each record is analyzed for applicability to any or all of the eight major files for which catalog cards are prepared. Each card file has its own criteria both in content and in the number and types of cards produced for it. The system produces a separate record for each card required, sorts together the records for each file, and alphabetizes within that file. Leading articles (regardless of language) are printed but are excluded in the sorting procedure. Cards are printed two-up in upper- and lowercase in the format prescribed by Anglo-American cataloging rules. After printing, the cards are distributed to the appropriate organizations and sections where they may be filed with a minimum of additional effort.

Monthly, a book catalog is compiled. This contains not only a listing by main entry but also indexes of personal authors, corporate authors, subjects, and titles. A biographic index (major personal author affiliation) capability is available although not presently used by NAL in the book catalog. This catalog is printed in varying numbers of columns changeable by control card option for each index. Again photocopy paper is used with a standard

upper- and lowercase (TN) print train. An alternate option is magnetic tape output formatted for direct input to a computer-driven LINOTRON. See bibliographic description for more detail.

Semiannually the index portions of the book catalog are cumulative. Main entry listings are not repeated. Multiyear accumulations may also be produced. The book catalogs are presently being published from photocopy printout by Rowman and Littlefield, Inc., New York.

Bibliographies, either scheduled or special, can be produced with the same indexes as those in the book catalog. These are normally prepared for printing via the LINOTRON. This magnetic tape record contains all formatting requirements with the exception of word divisions. Document title, page, and columnar (subject category) headers are provided by NAL. Running headers are inserted by the LINOTRON. Through predetermined codes, the CAIN tape specifies the print style, print size, and print format. Bibliographies may also be computer printed on photocopy paper similar to the book catalog.

Once a month, each record selected for publication is processed through a merge and adjustment program. At this point published records not previously on the permanent master file are added to it. Those which are already on it are compared and the resident record is adjusted to include the new user for whom the record has just been published. The term field is also verified and updated if necessary. Each term is also used to generate posting records for the subject authority file.

The permanent (published) CAIN data base is available on magnetic tape in either the master format or a print format of the linear proof (listing of each data element). Only records not previously published are added to the monthly sale tapes. These tapes may be ordered individually (new monthly selections) or collectively (whole file) at the cost of reproduction only. The tape is nine-track, 800 bpi, EBCDIC with standard IBM 360 header and trailer labels. One of the purchasers of CAIN tape is the CCM Information Corporation of New York which publishes *Bibliography of Agriculture* from it starting in 1970. Current purchasers include private corporations and universities, both in the United States and abroad.

The last type of output is normal computer printout of numerous internal reports in a variety of customized formats.

SEARCH SUBSYSTEM

The search capability of the CAIN system is not being used by NAL on its own data base at the present time. It is utilized, however, by other organizations who run the CAIN system on a parallel basis, maintaining their own data bases. The following description, therefore, pertains to the programmed system rather than to its use on the NAL data base.

This subsystem permits identification and retrieval of records in CAIN format based on search statements as applied to almost every data element

or combinations thereof. Such searches may use simple statements or a complex series of nested boolean parameters.

Questions may also be absolute or weighted to give more precise results. The weight factors if used are normally assigned to each statement within a search question, with a threshold weight assigned to the overall question. The total weight of all true statements must be equal to or greater than the threshold weight for the full query in order to be considered as meeting the search criteria. If such is not the case, the record will not be selected.

Since CAIN uses a controlled vocabulary, query statements on subject terms are first matched against that authority file. At this point each invalid (USE) term is replaced by a corresponding valid (UF) term if appropriate. In addition, if the query statement so specifies, the requested terms may be expanded one level in the hierarchy. In other words, it could generate additional statements requesting all broader, narrower, or related terms as specified if such structure were present for the subject within the vocabulary.

Because subject terms comprise the largest percentage of all search elements, an algorithm was developed whereby queries on this type of element are first processed against an inverted file. Identification numbers are extracted for all terms matching the query and only those candidate records are searched using the full query. On a serial file such as CAIN, this concept provides a substantial savings in computer run time.

The print options of retrieval output allow either for normal sequence by identification number or for a specific sequence as requested by the originator. The printout may contain all data elements or only those selected, all others being suppressed.

At the present time this subsystem is used infrequently by NAL and only for internal high priority searches due to the extremely limited subject indexing terms present. It is used more extensively on the parallel operation established for the International Tree Disease Register maintained for the U. S. Forest Service.

AUTHORITY FILES SUBSYSTEM

This subsystem updates, generates, expands, and maintains three types of authority files. These include subject terms with associated hierarchy, call numbers of indexed journals with abbreviated titles, and a subject term inverted file carrying the identification number of each record using that term.

Each transaction to add, change, or delete any data is both edited and reversed before entering the updating sequence. Thus an addition of a narrower term (for example, HORSE) to a base term (for example, ANIMAL) will automatically generate another transaction to add the broader term of ANIMAL to a base term (new or existing) of HORSE. This precludes having to manually enter both sides of an action as well as assuring reciprocity of entries. Due to the flexibility of the search sub-

system of CAIN, this hierarchical continuity is of great importance. If an item is changed the same procedure is followed.

In the instance of deletion, a broader precept is involved. In this case, the term is deleted from all entries in other hierarchies but is itself left on the authority file and marked as being no longer valid. It is thus available for search purposes but is not allowed to be used on subsequent CAIN data records.

During a normal CAIN data run, each call number or subject term in a record is verified against the appropriate file. Each element on these files is carried in two forms—one in stripped uppercase, and the other in preferred print form. When an incoming term is found on the authority file, the system substitutes the proper form. This includes substituting a valid term for an invalid term as in the "use—use for" relationship, as well as generation of the appropriate abbreviated journal title for a given call number.

In order to keep the authority file up to date, the transactions generated by the publication subsystem are now used to insert the record identification number into the inverted file as well as increase the number of postings per term. This assists search specialists in formulating queries in the manner which will reduce computer processing time to the greatest degree.

When published, the authority files themselves can be printed in a special format which displays the entire hierarchy of each term. In addition, up to ten levels of increasingly narrower terms can be listed for each term.

SUMMARY

CAIN is a broad-based comprehensive batch mode system which meets many library requirements. Its flexibility is apparent from the fact that it has already been expanded to select each newly cataloged serial record for transmission in MARC II communication format to the National Serials data bank being created by the three national libraries. Still more capabilities will undoubtedly be built into it before the NAL ultimate on-line system is implemented. The major thrust of the systems design has been to concentrate on simplifying user interface while imposing stringent and extensive service requirements on the computer system itself.

Due to its inherent fluidity, CAIN is being retained as an in-house system. It is so complex that a single change in one subsystem may have radial effects in any or all of the other portions. Continuing efforts are underway to simplify input, accelerate throughput, and expand its already generous services both to the staff of the National Agricultural Library and to those organizations utilizing output from the CAIN system.