INTER-LIBRARY SEARCHING THROUGH AN ON-LINE CATALOGUE

Harriet Velazquez
Eric Anttila
University of Toronto Library Automation Systems
130 St. George Street
Toronto, Ontario

ABSTRACT

Basic to the concept of resource sharing in libraries has been the production of union catalogues which describe the holdings of the participating institutions. The cataloguing processed through the facilities of University of Toronto Library Automation Systems (UTLAS) forms part of a database of over five million items. A recently introduced expansion to the on-line retrieval module of the Catalogue Support System (CATSS) now provides the potential to make this database accessible for use as a union catalogue. In order to accomplish this, several technical problems related to the large size of the database and its associated indices had to be solved.

RECHERCHE INTER-BIBLIOTHEQUE EN UTILISANT LES CATALOGUES COLLECTIFS

RESUME

La reproduction de catalogues collectifs offrant de différents périodiques est une nécessité pour un système de partage de ressources. Le catalogue reproduit par les systèmes d'automatisation de bibliothèque de l'Université de Toronto (UTLAS) comprend plus de cinq millions d'articles. Une expansion récente du système (CATSS) système de support du catalogue nous offre la possibilité d'avoir à notre disposition une base de données qui peut être utilisée comme un catalogue collectif. Afin d'accomplir cette expansion, plusieurs problèmes techniques relatif à l'indexation ont dû être résolvés.

THE DATABASE

University of Toronto Library Automation Systems (UTLAS) has been operating an on-line facility for shared cataloguing since 1973. Network libraries may derive cataloguing through the Catalogue Support System (CATSS) from any of the machine-readable files of four national agencies (Library of Congress, National Library of Canada, Bibliotheque Nationale du Quebec and the U.S. National Library of Medicine) or, after mutual agreement, with the other participating institutions. Bibliographic items new to the database may be contributed. Some 65 libraries or library consortia across Canada currently participate in the network. Each library builds a machine-readable file of its cataloguing, in MARC format, complete with local holdings including call numbers and copy and volume information. Extensive back-up and maintenance procedures ensure that each file contains the latest version of all records. The resulting database, composed of over five million "source" and "user" records is an up-to-date union catalogue of the holdings of over 500 individual Canadian libraries.

ON-LINE ACCESS

While the database has been accessible for sharing cataloguing data for six years, its capability for meeting other functions of a union catalogue -- inter-library loan, verification, and collection rationalization -- has been limited. This has been largely due to two conditions. The first, a restricted choice of access keys was the primary limitation. Access through an on-line index composed of Library of Congress Card Number, ISBN, ISSN and 40 characters of title was sufficient for cataloguing, where the system user had the item in hand. It was too awkward for reference type functions.

A further inconvenience was related to the economies which a library could effect by opting to store non-current records off-line. This meant that the ability to view the major portion of user files necessitated waiting for an over-night transfer to active storage. (Source files are kept in active storage.)

Enhancements to CATSS to make the union catalogue more accessible were made in two stages six months apart. In July 1979, the ability to search the database by authors, titles and subjects through a browsable, sequentially ordered index and by key word using Boolean operators was introduced. During the next several months, pilot projects were set up to test the new facility, and retrospective indexing of the databases of the institutions involved was undertaken. At the same time a method for creating shortened and compressed "mini-records" which would be continually available on-line was perfected. The indices, mini-records and an on-line interactive searching interface were inaugurated together in January as the beginning of Reference CATSS.

CREATING REFCATSS

The pilot projects chosen for testing REFCATSS involved indexing several files from the UTLAS database: UNICAT/TELECAT, a consortium of university, public and special libraries in Ontario and Quebec, having 670,000 records; CISTI, 35,000 records; Mississauga Public Library System, 140,000; and Canadiana, the National Library of Canada's MARC file of 75,000 records. Indexing produced 1,700,000 headings including all authors, titles and subjects. Keyword indexing was left to a second phase. Statistics, therefore, cannot be reported at this time.

In parallel with the indexing operation, mini-records extracted from every bibliographic record were entered into an on-line catalogue. Mini-records are subsets of the full MARC record and contain main entry, title, edition statement, imprint, collation and local call numbers and holdings.

The integrated author, title and subject index was not built directly record by record because of the processing time required. To minimize this time, index keys were generated containing the sortable form of the heading, truncated to 31 bytes, and compressed to an average of 4.8 bits per character by means of a variable bit string encoding technique which preserved sortability; the record sequence number or RSN, a system assigned number unique to each record; the MARC numeric tag and occurrence number of the field from which the heading was selected; and the size of the full sortable form of the heading before compression and truncation. These keys were created for every heading in every record in each of the databases. This resulted in 7 files of about 1,000,000 keys each including some duplicates because of restarts, etc. The 7 files were sorted individually and then merged into one big file of about 6,500,000 keys. A list building program then ran to eliminate duplicates and to create one record for each unique heading and its associated hit list. The output of this list building program was then the index. The 6,500,000 keys had been reduced in size to a file of 1,700,000 unique index entries. This resulted in an indexed sequential file consisting of three index levels and the data level. An access resulting in a direct hit requires four input/output operations; a browse type access or non-direct hit results in three input/output operations.

The mini catalogue record file, in order to save disk space, was compressed to an average of 5.1 bits per character while still preserving the full ALA character set. This compression was achieved by replacing each character with a bit string whose length is inversely proportional to the frequency of occurrence of the character in the database. While this technique means that CPU time is required in order to decompress the fields in the mini-record for display, the pilot project has shown that decompression amounts to an insignificant amount of time, about 2 1/2 minutes an hour.

The hit list for some index entries can get quite long; therefore

the arguments are kept in sort order to facilitate fast lookup when access to a specific file is sought. Record sequence number ranges are assigned to each database facilitating identification. As an additional measure, large hit lists are segmented into chunks 1024 hits long to allow segment by segment processing. Hit list display to the terminal is controlled by a user defined priority list of files to be accessed. This hierarchy of record sequence number ranges enables the program to control how many hits are actually displayed in a manner which makes most sense to the end user. For example, most users want to see hits in their own file first. These hits are made available and if the user is unsatisfied with one of them the next most important section of hits will be displayed; this could be the hits belonging to the closest institution geographically, if the user specified the display priority in that manner.

REFCATSS has the ability to offer the advantages provided by the controlled vocabulary of Library of Congress headings. An automated authorities system which verifies headings in bibliographic records filed through CATSS helps to ensure that a name or subject searched through REFCATSS will retrieve all the items related to that name. In addition, cross-references and see also references are indexed and a mini form of authority records is retrievable listing the authoritative form of the heading and the references linked to it.

The key word index is planned as a companion facility to the author title subject index. Key words will be extracted from all titles, corporate and conference names in bibliographic and authority records; and, from all subjects in authority records. The major problem which needed to be solved before implementing key word indexing was the handling of extremely large hit lists. The words 'United', 'States', and 'history' produce particularly large lists. The segmented hit list referred to above will allow these to be handled in a controlled way. Included in the hit list for keywords is the word number in the tag which provided the word. This feature will enable proximity searching. Users can then specify that a key word search be limited to those cases where another related word was used as well.

Boolean combinations have been made available using 'AND', 'OR', and 'NOT' operators. This will be extended to include modifiers taken from the mini catalogue record. For example, search for a certain word in publications since 1970; or, limit the search to films. Other planned extensions include the ability to save a hit list for later after some intermediate results have been processed; and, the ability to perform searches on user databases which have not been indexed in full themselves but which are accessible since the records were originally derived or copied from databases which are fully indexed. This latter function eliminates the need for repetitions on the hit list of bibliographic items which in fact are merely copies of each other.

IMPLICATIONS

An on-line union catalogue offers significant advantages over similar catalogues available through other media. The most important is that data can be available much more quickly. Time lags caused by filing backlogs or the publication of a new cumulation are not a problem. In addition, where replication and maintenance of large catalogues is normally financially prohibitive, an on-line catalogue resides and is up-to-date wherever a terminal exists.

Access to an on-line catalogue can be more flexible. Union catalogues are often limited to main entry access, requiring expensive manual verification before the catalogue can be used. Key words with Boolean operators can provide the capacity for identifying and locating items which could not otherwise be found.