the various arrays according to increasing value of the code numbers. The following sorted listings are printed: (a) six-digit apparatus code number, (b) tables (6) in which sorting is done for each individual code number, and (c) technique code numbers. In each table, the code numbers are printed followed by all the identifier numbers of those references in which that type of apparatus or technique is described (Figure 1b). As new references are added to the database (in alphabetical order), the program-assigned sequential identifier numbers will be changed. Thus one must always use the bibliography and numerical tables from the same database.

These tables allow one to locate quickly all references in which a particular type of instrument, device component, or technique is described. It is also possible to "design" an instrument by constructing a suitable six-digit number. The appropriate table can then be checked to see if a similar device has been described previously in the literature.

The program also contains an option to print the coauthors as cross-references. These names are contained on separate records (identified by a "3" in column 2) and are not assigned an identifier number by the program. Thus, even though they will be printed in the reference list, they will not appear as separate entries in the sorted tables.

Recently, a program has been written to reformat the database so that one complete citation is written per record. It is now possible to obtain selected bibliographies, in which the complete literature references are printed, for all articles dealing with any specific category. For example, one can list all articles in the database that were published in 1976, or all those describing a particular piece of apparatus, or some subset of components, or LTXRD technique.

ACKNOWLEDGMENT

This project was supported in part by Grant No. CHE 75-13935 from the National Science Foundation. The assistance of Frank Ely of the Adelphi University Computer Center and of Stuart Olson is gratefully acknowledged.

REFERENCES AND NOTES

(1) R. J. Feldmann, G. W. A. Milne, and S. R. Heller, "Crystallographic Data Retrieval and Display", *Trans. Am. Crystallogr. Assoc.*, **12**, 75–83 (1976).
(2) R. Rudman, "Low-Temperature X-Ray Diffraction: Apparatus and Techniques", Plenum Press, New York, N.Y., 1976.
(3) R. Rudman, "The Proper Description of Low-Temperature X-Ray Diffraction Apparatus", *J. Appl. Cryst.*, **10**, 209–210 (1977).

CASSI and the Compression of Journal Names in an Information Retrieval System

J. HEILIK and R. H. BURTON*

Canada Institute for Scientific and Technical Information, Ottawa, Canada K1A OS2

Received June 21, 1977

A data compression technique for journal names is described. This technique is enhanced by use of the Chemical Abstracts Service Source Index (CASSI). Implications of CASSI's use in an information retrieval system are then discussed.

The Canada Institute for Scientific and Technical Information (CISTI) began offering an on-line information retrieval system (CAN/OLE) on a national scale in April 1974. Experience with the system quickly indicated several potential bottlenecks which could restrict the usefulness of the system and/or increase its costs. Of critical economic importance was the one of limited availability of disk storage space for data.

CISTI buys computing services from the National Research Council (Canada) Computation Centre. With file sizes soaring into the millions of records, the disk space at the Computation Centre was quickly filled, and additional disk drives, dedicated to the sole use of CAN/OLE, had to be acquired. These devices are expensive, so priority was placed on keeping the requirement for them as low as possible. One way is to find methods of compressing data. Several techniques were found suitable and are used currently; the subject of this paper is the one dealing with journal names.

A journal may have several articles per issue, several issues per volume, and several volumes referred to in OLE. Each article gives rise to an OLE record; each record contains a journal name as part of its bibliographic citation. Consequently, all records coming from a given journal will have identical data in the journal name field. Early in the life of CAN/OLE, it became evident that although the files contained many records, the number of journals was relatively small. Hence, this duplication of data became fairly significant.

Based on the above observation, a simple compression technique was developed. All journal names were removed from the base files, and replaced with a five-character CODEN (the standard ASTM CODEN stripped of its check digit). A central file of journal names was created in CODEN order. Whenever a record is displayed at a terminal, a name is automatically fetched from the journal file and displayed with that record. To the person at the terminal, it appears that all records have journal names. The link between records and the journal name file is, of course, the CODEN.

Removing journal names reduced storage requirements by 13%. In May 1977, there were approximately 4.5 million records in the on-line files; this modest 13% is thus equivalent to 585 000 records for which no additional space is needed. Evidently, when very large numbers are involved, even the smallest gains are significant.

A common problem faced by a customer of CAN/OLE, or most other bibliographic information retrieval services, is that having found a desirable set of references, how does he find the corresponding documents? Creating a central journal name file provided CISTI staff with an opportunity of alleviating this problem somewhat. Being relatively small and easily controlled, the name file was amenable to experimentation.

In mid-1975, CISTI approached Chemical Abstracts Service (CAS) to suggest an experiment with the Chemical Abstracts Service Source Index (CASSI). This Index was of interest because it contains not only unabbreviated journal names, but a record of which libraries subscribe to the various journals. Further, it contains all journal names found in two of the CAN/OLE data bases (*Chemical Abstracts Condensates* and

```
Despiay record before compression
     LUCERNE-D ESTABLISHMENT STUDIES ON UNCULTIVATED COUNTRY PART I
     GERMINATION AND SEEDLING ESTABLISHMENT
     JANSON C G; WHITE J G H
     NZJ AGRIC RES
     1971, 14 (3).; 572-586.

Despiay record after compression (using CASSI)
     LUCERNE-D ESTABLISHMENT STUDIES ON UNCULTIVATED COUNTRY PART I
     GERMINATION AND SEEDLING ESTABLISHMENT
     JANSON C G ; WHITE J G H
     NEW ZEALAND JOURNAL OF AGRICULTURAL RESEARCH (FORMERLY NZTAA)-CISTI
     1971, 14 (3).; 572-586

Despiay record before compression
     TEMPERATURE EFFECTS ON CHLORELLA-PYRENOIDOSA PHOTOSYNTHESIS IN LARGE
     SCALE CULTIVATION IN OPEN BASINS
     BERDYKULOV K A
     UZB BIOL ZH
     1970, 14 (1).; 20-22

Despiay record after compression (using CASSI)
     TEMPERATURE EFFECTS ON CHLORELLA-PYRENOIDOSA PHOTOSYNTHESIS IN LARGE
     SCALE CULTIVATION IN OPEN BASINS
     BERDYKULOV KA
     UZBEKSKII BIOLOGICHESKII ZHURNAL (FORMERLY TAUBA)
     1970, 14 (1).; 20-22
```

Figure 1. Before and after examples of journal information.

BIOSIS), and there were possibilities that it would, sometime in the future, include those of two others, *Engineering Index* and INSPEC. The CAS response to the suggestion was most favorable, and CISTI subsequently received a license to use CASSI for one year at no cost other than that of tape reels and mailing.

The project proposed was to replace the journal name file with CASSI. The OLE customer would receive two benefits from this replacement. The first was that in records retrieved from *CA Condensates* or BIOSIS, he would be given full, unabbreviated journal names. (On occasion, abbreviations can be rather obscure.) The second was that he could know immediately whether or not the document corresponding to a given OLE record was available from CISTI.

After careful study, the following subset of the CASSI record was chosen to be the OLE journal record:

1. Tag 55 — CODEN. Used as the key for the journal file, and to keep all data pertaining to one journal title together
2. Tag B2 — Former CODEN. Used to add the text "(FORMERLY coden)" to the end of a journal name if required
3. Tag 9F — Type. Used to set up a note indicating one of "(CEASED)" if the journal has ceased publication or "(FORMERLY. . .)" if it has undergone a name change
4. Tag A2 — Key title. Used to provide the customer with an unabbreviated journal name.
5. Tag F2 — Successor CODEN. Used to set up the note "(CONTINUES AS coden)" if applicable.
6. Tag 115 — Canadian holdings. CASSI, being a union list, is searched for the CISTI location symbol. If found, the note "(CISTI)" is appended to the journal name.

Because CASSI comes in Standard Distribution Format, only minor modifications to CISTI software were required to convert the above six data elements to the journal name file format, this being indexed sequential, with the CODEN serving as the index key. The converted subset of the CASSI file was then compared with the OLE journal name file; when a match was found, the CASSI record bumped the original. The result was a new journal name file, most of whose records came from CASSI.

Maintaining this name file is a two-step operation. As the OLE base files are updated, quite often new journals not yet in the name file are found. These are added "as is" (usually abbreviated) to the name file; the base file record will contain only the CODEN as usual. Every calendar quarter, CAS issues an update to CASSI. When received, the CASSI update is compared with the name file. As with the original file creation, any match found implies the new CASSI record will replace what is already there. Totally new CASSI records are simply added to the name file. The matching of records is done using the CODEN field.

The system would be perfect if every record in the base files had a unique CODEN corresponding to one in CASSI. Unfortunately, this is not the case. In the original matching of the journal name file with CASSI, CODEN were found in the former which were not in the latter. Further, a number of records have CODEN-like codes rather than actual CODEN. All books have the common code BOOKA. Codes for patents simply group patents by country. Abstracts picked up by BIOSIS from other abstracting services (e.g., *Referativny Zhurnal*) carry the code of the abstracting service, not of the article being abstracted. The above explains why the compression technique was restricted to journal names, and not used for book names, patents, and, for similar reasons, conference proceedings and technical reports.

For records in which these anomalies are detected, the name file is not used. Instead, the original bibliographic information is kept unaltered in the base file records, and appropriate flags are set indicating to OLE that it need not go to the name file. The number of these problematic CODEN and codes is in the order of 2500, or about 6% of the total.

The initial impetus for the name file project was the saving of space; in that we have seen it was successful. With the aid of CASSI, it was successful on a different level as well; it enabled the OLE customer to get better journal information at less storage cost. It further helped him find the actual article by informing him of its availability at CISTI. Consider the before and after examples in Figure 1.

In fact, the marking of CISTI owned journals is but one step in a plan to link the entire CISTI collection to its information retrieval services. Recently made available in Ottawa is a facility by which documents can automatically be requested from CISTI through OLE. It is worth noting that CISTI, because of its extensive collections, is Canada's largest supplier of scientific and technical documents in the national Inter-Library Loan network. For the OLE customer to be able to tap directly into this document source greatly enhances the information retrieval service. Future developments will further strengthen this link. We are most grateful to CAS for its enthusiastic cooperation and useful advice in launching this development work.

REFERENCES AND NOTES

(1) J. Heilik, "CAN/OLE: A Technical Description", Canadian Association for Information Science, Fourth Canadian Conference on Information Science, London (Ontario), May 11–14, 1976, Proceedings, pp 47–55.
(2) "Continuing Innovation in Information Systems – CISTI Opens Its Doors", Science Dimension (National Research Council Canada), Vol. 6, No. 6, 1974, pp 4–7.
(3) American Chemical Society, Chemical Abstracts Service, Chemical Abstracts Service Source Index, Columbus, Ohio (Cover title: CASSI; Chemical Abstracts Service Source Index).
(4) American Chemical Society, Chemical Abstracts Service, Specifications Manual for Computer-Readable Files in Standard Distribution Format (International Standard Book Number: 8412-0150-1), 1977.