A DECENTRALIZED, COOPERATIVE, INDEXING PROJECT: CANADIAN INDEX TO GEOSCIENCE DATA (UN PROJET COOPERATIF ET DECENTRALISE D'INDEXATION)

Katherine L. Gunn and C.F. Burk, Jr.
Canada Centre for Geoscience Data
Ottawa, Ontario, K1A 0E4

ABSTRACT

A project to identify sources of geoscience data (i.e. original observations and measurements) was begun in 1967 by the federal Department of Energy, Mines and Resources. Because of the large number of documents involved and their wide physical distribution a decentralized, cooperative approach was adopted. Canadian geoscience agencies were asked to voluntarily index documents published or held by each agency, the results being incorporated by the Canada Centre for Geoscience Data into a computer-based index. This approach has had valuable results not likely possible through centralized bureaucratic means. During the past 8 years, over 40,000 titles from 10 agencies have been indexed in detail, and consolidated into the Canadian Index to Geoscience Data. Consistency and vocabulary control are maintained through a Thesaurus under direction of a committee of the contributing agencies. Good coverage now exists for some areas of Canada, notably Ontario, Saskatchewan, Newfoundland, Quebec, Yukon and Northwest Territories, with general completion of federal and provincial government documents expected within about four years. (Une projet d'identification des sources de données géoscientifiques (i.e. observations et mesures originales) a été initié en 1967 par le ministère fédéral de l'Energie, Mines et des Ressources. A cause du grand nombre de documents impliqués et de leur distribution physique étendue, une approche coopérative et décentralisée fut adoptée. On demanda aux agences géoscientifiques canadiennes d'indexer, sur une base volontaire, les documents publiés par, ou en possession de chaque agence, les résultats étant incorporés par le centre canadien de données géoscientifiques dans un fichier index informatique. Cette approche produisit des résultats valables, qui n'auraient probablement pas été obtenus par des moyens bureaucratiques centralisés. Au cours des 8 dernierès années, plus de 40,000 titres de 10 agences ont été indexés en détail et incorporés à l'index canadien des données géoscientifiques. La consistence et le controle du vocabulaire sont maintenus par un "Thesaurus" sous la direction d'un comité constitué des agences participantes. Une bonne "couverture" existe maintenant pour quelques régions canadiennes, notamment l'Ontario, la Saskatchewan, Terre-Neuve, le Québec, le Yukon et les Territoires du Nord-Ouest. Le tout devrait être complétés par les documents des gouvernements fédéral et provinciaux d'ici environ quatre ans.)

A DECENTRALIZED INDEXING PROJECT

INTRODUCTION

During the early 1960's the geoscience community in Canada shared a common problem with many other disciplines, realizing that the quantity and complexity of available information was rapidly becoming unmanageable, and asking whether systematic compilation and storage, mainly by computer, would help. How much effort would be necessary to control the expanding quantities of data? Could control be exerted quickly enough to make existing data useful and plan for the future before everyone became buried? What could be done?

Senior representatives of provincial and federal governments, academia and industry, who were members of the National Advisory Committee on Research in the Geological Sciences (NACRGS) turned their attention to the situation in 1965 by appointing the ad hoc Committee on Storage and Retrieval of Geological Data in Canada. A sequence of gradually more detailed studies by the latter Committee led to a set of recommendations which were presented to, and accepted by, the Conference of Provincial Ministers of Mines and NACRGS in 1967 and 1968. Action on these recommendations led to the formation of an office in the Geological Survey of Canada, Department of Energy, Mines and Resources, which in 1970 became the Canadian Centre for Geoscience Data. In April 1974 the Centre became an independent division with the office of the Assistant Deputy Minister, Science and Technology, of EMR.

The areas examined by the ad hoc Committee were largely those of 'hard' data - measureable, reproducible facts such as chemical analyses, mineral production, fossil names, etc. It will come as no surprise to hear, however, that early in its studies the Committee decided that sources of data - reports, maps, files- in short, documents of all kinds- must be located and described before the data in them can be used in a systematic, let alone complete manner. Thus one of the major recommendations of the Committee was that an index be established on a national basis which would

> "...include reference to all published data and any unpublished data that might be made available voluntarily in Canada. The Index would not contain the actual geological data or even an abstract, but would indicate what kind of data is available and the sources from which it could be obtained. ...The focus of emphasis on data, including unpublished data, creates a National Index that is not conceived of as competing with any of the bibliographic indexes currently being produced in North America ... . It was further concluded that a National Index to geological data could exist separately from, and could be proceeded with in advance of ... storage and retrieval of geological data." (Brisbin and Ediger, 1967, P. 23)

A DECENTRALIZED INDEXING PROJECT

To carry out this recommendation, the Canadian Index to Geoscience Data was initiated as a pilot study in 1966, and in 1967 became a formal project of the Geological Survey. By 1969, three provincial departments of mines and two federal departments holding earth science documents had voluntarily begun indexing.

WHY 'DATA'? - WHAT AND WHERE

Generally,a documentation file or secondary source is set up to catalog or index a specific collection of documents, or at least all documents touching on a given subject. The Canadian Index was, as far as we know, unique at its inception in limiting its coverage to those documents which contain 'hard' data. This is probably the inevitable result of the fact that, in the deliberations described above, the ad hoc Committee arrived at the need for an index in the first place from a concern about the management of such data, not from an interest in the documents per se. The same requirement has been recognized by the Committee on Data for Science and Technology of the International Council of Scientific Unions, as described in its report on "Energy Data, Accessing and/or Retrieval" (CODATA, 1974).

A precise definition of 'data', and a distinction between data and interpretations, presented difficulties to the ad hoc Committee, and has remained an elusive semantic and conceptual problem. (It would now appear that the difficulties arose because data are not inherently different from interpretations -- one person in one context observes and records some facts (data), and draws from them conclusions; a second person in another context may accept these conclusions as facts (data) leading toward his conclusions. Many people retain an instinctive feeling that some pieces of information are more 'factual' than others, but there has been a realization that the boundaries depend on context rather than absolute definition.) In addition to only admitting documents which contained data, the Index provided for 'tagging' the data by an indicator attached to appropriate keywords. Thus a report describing a mining area might have keywords for COPPER and NICKEL which occur in that area, and keywords for GOLD and SILVER with 'D' weights attached because some numerical data on these two latter commodities were included in the report.

The definition of documents to be included in the Index has recently been broadened from that described above, to include those documents which are part of a series held or published by a contributing agency and of which the majority are suitable for indexing. Thus the contributing agencies can rely on the Index to provide complete catalogs of their series of reports.

A DECENTRALIZED INDEXING PROJECT

CONTRIBUTING AGENCIES

The National Advisory Committee on Research in the Geological Sciences and the Mines Ministers' Conference urged the organizations represented in their ranks to voluntarily contribute the indexing effort necessary to construct the Canadian Index. Several provincial departments of mines began indexing in 1968, in addition to two federal government offices. The participation has gradually increased as the utility of the Index became more apparent. As of early 1975, all but two provinces had committed staff and resources necessary to the project, as well as five federal branches in two departments. One mining company has indexed and contributed its exploration assessment reports.

The decentralization of the indexing effort has produced benefits in accord with each of the reasons advanced by the ad hoc Committee for decentralization. The wide physical distribution of the documents, frequently in only one existing copy, would have made indexing most difficult for a central agency, but the concurrent work of up to ten indexing offices has greatly speeded growth of the file. The agency indexers know the documents, and their uses, better than a small generalized staff ever could. Each agency has one or two officers to act as effective ambassadors for the Index in their organization, and as spokesmen for their agency to the Canada Centre in the management of the Index.

Potential contributors to the Index include the remaining provinces, other mineral and petroleum exploration firms, professional societies and research organizations. If and when additional funding becomes available, the Centre may be able to support indexing of the general scientific literature and university theses.

DOCUMENTS

The projected coverage by the Canadian Index to Geoscience Data encompassess all documents (reports, maps, files) which are sources of geoscience data, dealing with Canadian territory, and if unpublished, are made available for indexing.

The size of the Canadian Index in early 1975 is summarized in Table I. As has consistently been the case, over half the data sources recorded in the Index are unpublished - single, or at most triple, copies of mineral exploration evaluation reports, government open files, etc., whose existence would otherwise often be recorded only internally or at best on limited distribution lists. A word on confidentiality is appropriate here: although a document itself may be confidential, the title, reference and keywords are often not so considered.

A DECENTRALIZED INDEXING PROJECT

Geographic coverage of Canada is essentially complete for the federal agencies whose purview is the whole country, as well as for six jurisdictions at the provincial/territorial level (see Table I). It is estimated that another four years' effort will result in good coverage for all provincial and federal agencies.

INDEXING

The indexing of geoscience documents is carried out on site by contributing agencies, in most cases by a permanent staff member (who may or may not have other duties) and in all cases by professional geologists. Results of the indexing, in the form of computer-readable input, are transmitted to the Canada Centre for Geoscience Data in various forms (punched cards, magnetic tape, telephone communication lines), where they are consolidated for input to the master file at periodic intervals.

One of the distinctive and most valuable features of the Index lies in the systematic geographic control which is available for all documents. This need results from the fact that geoscience deals with rocks, soil, mineral resources, etc. which are spatially distributed, and retrieval of these data is almost invariably couched in geographic terms. The National Topographic System (NTS) provides a national grid within which geographic locations can be recorded in the Index to within about 40 miles. The NTS grid is represented by seven-character codes which are as familiar to geologists and geographers as their home street numbers!

MANAGEMENT OF THE INDEX

The day-to-day operation of the Index is conducted at the Canada Centre by the Index manager, a geologist, who on occasion has had a junior geologist as assistant. Housekeeping steps include review and coordination of the input and indexing procedures of the agencies, computer runs to update and edit the file, and searches to prepare the customer products described below. Longer range considerations involve technical improvements in all aspects of the project, promotion and dissemination of the products of the Index, and development of new sources of input.

The contributing agencies maintain a voice in management of the Index through representation on the Geoscience Index Advisory Committee, whose terms of reference may be summarized as follows:

1. To advise the Canada Centre for Geoscience Data on the most effective means for disseminating information contained in the Canadian Index, including services and products of particular benefit to the contributing agencies themselves. Of concern here are such matters as the physical form of indexes, application of new technology such as microform and on-line access, and liaison between CCGD and the contributing agencies.

2. To provide a forum in which workable and satisfactory operational procedures can be established for CCGD and contributing agencies to the Canadian Index to Geoscience Data. These matters relate to the details of indexing, including vocabulary control, data entry, and transmission of this data to the Canada Centre for Geoscience Data. An Indexing Manual and Thesaurus form the core of these activities.

A THESAURUS - WHY BOTHER?

Whether or not to control the vocabulary in a secondary file will probably always be a moot point, with strong convictions or bias on both sides. There are four main reasons for maintaining a Thesaurus throughout development of the Canada Index to Geoscience Data -

1. The geosciences, as a amalgam of many disciplines, from chemistry to zoology, are rife with terminological problems, and without some semblance of control and highlighting of relationships, the use of many keywords would be at least difficult, if not misleading.

2. The Index began its life in a relatively primitive, batch-oriented computer environment, in which a thesaurus provided a ready overview of the file contents, which could otherwise only be reviewed in costly and time-consuming listings.

3. Consistency of approach in indexing for a decentralized operation requires a thesaurus.

4. In a file limited to essentially one subject area, controlled vocabulary can provide the user with a high assurance of finding what he needs.

A DECENTRALIZED INDEXING PROJECT

The Canadian Index thesaurus initially was permitted to grow out of author terminology - almost not a thesaurus! As the file grew in size and coverage, patterns and frequencies in usage were noted, and guidelines drawn up to provide some uniformity of intellectual approach. Precise rules apply to the format of keywords, and both format and previously authorized content are controlled by the computer software, as described below.

The thesaurus recently has been subdivided by broad subject areas - not classified - for the convenience of both indexer and user, a step which has been welcomed on all sides.

COMPUTER SOFTWARE

The Index file is operated on the Energy, Mines and Resources Control Data CYBER 74 computer. Current software is a decendant of the Streamed Information System (SIS) of Imperial Oil Limited, which was provided to the federal government for use by the Index in 1966. A second version, SIS II, was developed into the present system, RAID (Reference AID), by the Computer Centres of EMR, the University of Calgary and the University of Saskatchewan.

RAID contains modules for updating, editing, thesaurus control, retrieval, sorting, printing and statistical listings. Editing and error reporting are especially detailed, and although these make the system somewhat expensive to use, they obviate much human effort in checking and correcting. Keywords are not accepted in reference to titles unless they have been authorized into the thesaurus, and most other fields are checked for adequacy and/or accuracy. The resulting error reports are checked centrally, and returned to the indexers for review.

The system operates only in batch mode, but with the CYBER 74 being accessible through timesharing terminals, all batch work can be run from the Canada Centre on terminals such as VUCOM I, DATACOM 300, and TELETERM 1132. RAID is currently used in parallel by one agency to operate a subset of the Index on its own IBM 360, and by another to access subfiles of its own documents on the EMR CYBER 74.

The file structure generated by RAID is sufficiently straightforeward that little difficulty has been encountered in operating the file under other systems or on other installations. Extraction of certain data elements and conversion to a completely different type of file has been accomplished with somewhat greater effort.

A DECENTRALIZED INDEXING PROJECT

PRODUCTS

Access to the Canadian Index to-date has been through the Canada Centre for Geoscience Data. Retrievals from the file are available as printout, magnetic tape or microfiche, at the user's request. The updating frequency of the file is sufficiently long - 2 to 6 months - that searches are all retrospective. The products of the file fall into three categories:

1. General Indexes — Complete indexes by province; released as a group to cover the whole country; highest frequency is annual.

2. Special Subject Indexes — Of more limited scope, covering one or several provinces for a particular subject such as a mineral commodity; released by the Centre as public interest would appear to warrant.

3. Custom Indexes — One-off searches on request for any user; distributed only to that customer.

In addition to release by the Canada Centre, all the above types of indexes may be prepared for contributing agencies to distribute as publications or catalogs.

The indexes produced by RAID may list only title-citations, or may list associated keywords as well with each title, and may be sorted to three levels. The presence of an index (or accession) number permits manual co-ordinate searching between keywords. The statistical reports produced by RAID, such as keyword frequency, permit remarkably accurate estimates of the potential number of hits for a given search, once some familiarity has been gained with the file.

COSTS

The cost of preparing input to the Index, from review of a document by the geologist indexer through transmission to CCGD, ranges from $2 to $13. The lower cost covers a two-page item which has two keywords and the higher may be a memoir of 300 pages with several large maps.

Updating of the file under RAID, at the current size of 40,000 documents, costs about $6,000-$10,000 annually, covering a variety of computer runs which can be subdivided roughly into 2 to 4 complete updates. This amount includes assorted other housekeeping steps - master listings, edit reports, etc.

Searching of the file is highly variable in RAID, with costs ranging from $20 for access to a sub-set of the Index to $400 for a large number of hits (say 3,000) sorted in an intricate manner. A typical range is $80 to $200, for Custom Indexes.

THE FUTURE

The future of most secondary files probably lies in interactive access. The increased speed of access is attractive to users, and the flexibility of searching opens new dimensions in user involvement and understanding. We are examining the characteristics of our file, the requirements of our users as we understand them, and the availability of interactive systems to plan our move into this area.

A dispersed data collection project such as the Canadian Index is bound to suffer confusion in the handling and transmission of input from a dozen offices to the Centre. We have found the situation has improved markedly with each automated step that has been introduced - transmission of magnetic tapes instead of cards; on-line input and editing of data instead of hand-printing and keypunching - so that we anticipate even smoother operations with further advances such as optical character recognition of input, off-line accumulation of data for telephone-line transmission and data entry on intelligent terminals, which are currently being planned or investigated.

A new area of indexing, although a logical outgrowth of both the mandate of the ad hoc Committee and the methods of the Index, is with computer-based data files. One pilot project (MINDEX, Gunn, 197*) has been well received which identified and described Canadian computer-based files that store the 'hard' data concerning mineral and fuel deposits.

BIBLIOGRAPHY

BRISBIN, W.C. and EDIGER, N.M., 1967 Editors, A National System for Storage and Retrieval of Geological Data in Canada. National Advisory Committee on Research in the Geological Sciences. Available from Geological Survey of Canada, Ottawa 175 p.

BURK, C.F. Jr. 1969 Supply and demand of Geoscience data. Western Miner, 42(2): 30-36

BURK, C.F. Jr. 1969 Use of computers find increasing role as effective tool in minerals search. Northern Miner, 55(36): 69-71

BURK, C.F. Jr. 1970 Geoscience data index available. Oilweek, 20(51): 10-11

* In press

A DECENTRALIZED INDEXING PROJECT

CODATA Working Panel on Data Tagging 1974 Energy Data - Accessing and/or retrieval. In CODATA Bulletin 12, Paris, France.

GUNN, K.L. in press MINDEX: An index to mineral deposits data files in Canada. Geological Survey of Canada Paper Series.

MCGEE, B.A. 1972 New Key to mineral exploration: Canadian Index to Geoscience Data. Canadian Min. Journal, 93(4):43-47

MCGEE, B.A. and BURK, C.F. Jr. 1972 The Canadian Index to Geoscience Data: a new national service for mineral exploration. Proc. Data Proc. Inst. Conference 70, Ottawa, P. 115-122

MCGEE, B.A. 1969 The Canadian Index to Geoscience Data. in Report of Activities, Part B: Nov.1968 to Mar.1969. Geological Survey of Canada Paper 69-1, pt.B,49

TABLE 1

CANADIAN INDEX TO GEOSCIENCE DATA
A SUMMARY - FEBRUARY 1975

Operated by - Canada Centre for Geoscience Data
Department of Energy Mines And Resources
Ottawa, Ontario, Canada

Began - 1967

Size - 40,916 titles
- 225,000 keyword occurrences
- 5000 geological descriptor keywords

CONTRIBUTING AGENCIES	jurisdiction	coverage to 1975
Dept. of Mines and Energy, Newfoundland	Nfld.	partial
Dept. of Industry and Commerce, P.E.I.	PEI	begin 1975
Dept. of Mines, Nova Scotia	NS	begin 1975
Dept. of Natural Resources, Quebec	Que.	partial
Division of Mines, Ontario	Ont.	complete
Dept.of Mineral Resources, Saskatchewan	Sask.	complete
Alberta Research Council	Alta.	begin 1975
Dept.of Mines & Petroleum Resources, B.C.	B.C.	begin 1975
Dept. of Energy, Mines and Resources -		
Geological Survey Branch	Canada	partial
Earth Physics Branch	Canada	partial
Mineral Development Sector	Canada	complete
Dept.of Indian and Northern Affairs-		
Exploration and Geological Services	NWT/YT	complete
Oil and Gas Lands Exploration	NWT/YT	partial
Churchill Falls Labrador Corp. (BRINEX)	corporate	to 1973