TRANSDAC
AN INFORMATION STORAGE AND RETRIEVAL SYSTEM

J.K. Hendry
Canadian Transport Commission
Ottawa, Ontario. K1A ON9

ABSTRACT

Transportation regulation and research necessarily interact with or impact on many other industries, disciplines, and concepts. In order to establish and maintain effective access to competent sources in the spectrum of information resources, a computer-based bibliographic storage and retrieval system has been established in the Canadian Transport Commission. This system, emphasizing quality of documentation rather than quantity, provides rapid access to an international inter-disciplinary database of unique or continuing value to strategic, managerial, and operational functions. All reference materials are heavily indexed and most contain detailed descriptive commentary on their scope, focus, and principal content. A semi-conversational user-system interface permits non-technical personnel to directly and independently retrieve materials from the database using a remote-access time shared terminal system. (Les réglements et les recherches en transports réagisaent reciproquement et nécessairement avec beaucoup d'autres industries, disciplines et concepts. Pour établir et maintenir un acces efficace à des sources compétentes dans le spectre des ressources de renseigements, un système bibliographique d'accumulation et recouvrement basé sur un ordinateur a été établi dans la Commission Canadienne des Transports. Ce système, accentuant la qualité de documentation plutôt que la quantite, fait provision pour un acces rapide a une donnée internationale et interdisciplinaire de valeur unique et continuante a des fonctions stratégiques, directonales et en pleine activité. Tous materiaux de reférénce sont bien dressés et la plupart contiennent un commentaire descriptif et détaillé sur leur portée leur sujet d' interet et leur contenu principal. Un entrelacement qui est etabli pour un utiliseur du système et qui est en language automatique

TRANSDAC

permet au personnel de recouvrer, directement et indépendament, les materiaux de la donnée en faisant usage d'un systéme terminal qui a un acces à distance et qui est un travail en parallele.

INTRODUCTION

TRANSDAC is a computer-based information storage and retrieval system which has been operating in the Transport Industries Group of the Canadian Transport Commission for two years.

A principal advantage of the system is its application to most text-processing needs of small groups, or of particular individuals, within larger organizations. It permits an unlimited number of users to share a common software base and, at the same time, to independently create and manipulate computer files reflecting the unique information requirements of each.

Although a number of independent information files are operating now, including a database maintained by the Canadian Transport Commission Library, this paper will focus on development of the TRANSDAC system in response to a specific information management problem. Other applications, of which over one hundred have been identified, are readily accommodated.

Because the system operates essentially on an informal basis with a staff of one, it cannot and does not attempt to function as a conventional library information referral service for other than the immediate internal user group. Steps are being taken, however, to provide other responsibility centers in the Canadian Transport Commission with "own-account" access to either or both the software and databases.

GENESIS

The initiative to develop this system originated with the author's problems in managing a personal reference collection which, by late 1971, comprised 500 citations to documents of unique or continuing value to a wide range of socioeconomic and industrial research programs. Since further extension of the already complex manual indexing system was clearly inadequate, it was decided to establish a computer-based system, subject to the following criteria;

(1) the system must come "off the shelf", requiring little or no program development

(2) it must have minimal acquisition cost

(3) it must provide a reasonably non-technical interface between software systems and the user

(4) it must have flexible text processing capabilities, and

(5) it must be economically and operationally feasible at low levels of use

With these criteria as benchmarks it was then quite simple, using the existing reference base, to locate and review a large number of computer oriented systems, the most promising of which was the FAMULUS system in use in the United States Department of Agriculture: (Yerke, 1969). With the generous assistance of Mr. James Clarke in the U.S. D.A., a copy of the program was obtained by mid-1972. During the latter part of the year, all software compiling, including debugging, program modifications, and insertion of new programs was completed by Dr. Richard Lee on the IBM 360/85 installation at Systems Dimensions Ltd. in Ottawa.

Simultaneously, a large number of experimental coding/indexing formats were tested, followed by a gradual "loading" of the 500 document database. A further 1000 citations have been added since, on a highly selective basis, out of an estimated 50,000 titles reviewed. Experience suggests that the threshold limit on this file will be about 2000 references, constrained not by system capability but by the selection criteria. Materials of a transitory or more topical nature are being accumulated on a separate file as time permits.

INDEXING AND RETRIEVAL

Information on each document is distilled into eight citation fields;

TITL	official title
AUTH	personal and/or corporate author
ABST	abstract
PUBL	publisher
SUPP	date, length, cost, frequency
AVAL	availability of copies in Ottawa
TYPE	classification of originating source
DESC	descriptors/keywords/key phrases

TRANSDAC

With user comprehension as a first priority, abbreviations, short-forms, or exotic coding methods, which admittedly save time and storage costs, are specifically excluded. Under this format, the file can be searched on the basis of any permutation of title, author, date, subject, geopolitical origin, or other criteria. Usually only the DESC field is scanned, although any or all of the others may be selected.

In order to supplement standard searches and optimize use of data processing resources, the file structure is also designed to permit special bibliographic compilations;

(a) on a particular subject in a particular library
(b) of materials published by any or all levels of government
(c) of all daily, or monthly, or annual publications
(d) of only statistical, directory, policy, instructional, etc. materials.

These and other combinations are used frequently with notable effectiveness.

User-oriented data, in addition to conventional bibliographic information, include descriptive abstracts up to 300 words in length and notations on the location of many of the documents in Ottawa. (Only 15 per cent of indexed materials are held in the Canadian Transport Commission Library).

The composition of database resources, in terms of each document's specific orientation, is approximately as follows;

Transportation	30%
Other industries (mining, agriculture,...)	40%
Disciplines (economics, social sciences,...)	15%
Concepts (technology, inflation,...)	15%

Particular citations can refer to books, manuals, maps, films, monographs, journal articles, serial publications, statistical reports, or any other information media. Input data are obtained from the publications directly, library accession lists, publishers' announcements, abstract journals, and many other sources. Although every effort is made to ensure database integrity, the frequent use of unofficial citation sources presents a continuing problem of accuracy and

completeness. Given existing constraints however, bibliographic "purity" has been temporarily sacrificed in favour of recording the essence of many information resources.

TECHNICAL SPECIFICATIONS

The TRANSDAC software package is comprised of 8 main program modules and 55 subroutines written in FORTRAN IV G-level for the IBM 360 series. It can be made compatible with certain Univac and Control Data Corporation equipment, but is presently installed on an IBM 360/85-370/158 configuration at a commercial service bureau (Systems Dimensions Ltd.) in Ottawa.

All software and some datasets are maintained on direct access devices (disc packs), with most information processing activities conducted over remote-access terminals from C.T.C. premises.

The system provides for variable length records (citations) up to 4000 characters in length with up to 10 separate fields of information in each. Information retrievals are conducted using both standard Boolean operators (and, or, and not) and, in a properly structured file, the relational connectors less than, equal to, not equal to, and greater than. In practice, retrieval permutations are limited only by the ingenuity of file structuring defined by the user and can include numerical and textual data, the latter consisting of words, phrases, or word fragments. Any field or all fields can be searched at one time, and searches can be conducted individually or in batches.

The system provides also for internally generated thesauri and vocabulary listings which are alphabetically and numerically indexed to each master information file.

OPERATING COSTS

The costs detailed below for principal TRANSDAC operations are based on commercial service bureau rates in 1974 providing online access to a bibliographic database totalling 1.5 million alphameric characters, representing 1208 heavily indexed documents with detailed abstracts. These costs may be readily discussed in terms of low frequency and high frequency operations, the latter relating specifically to file searching and information retrieval. Low frequency operations include database creation, updating, sorting, merging, and indexing.

Job costs in the latter category range from a high of 6.50 cents per record (citation) processed to a low of about .40 cents. The highest unit costs are associated only with file creation using the EDIT program and reflect primarily the expense of reading punched card data and copying it to tape. Subsequent operations rarely exceed unit costs of 1.5 cents per record. The only exception to this occurs when the GALLEY program is used to print selected records, in which case all computer processing and overhead costs may be allocated to as few as 2 or 3 items. When GALLEY is used to print an entire file, unit costs average 2.80 cents per record, with copies in triplicate. File sorting (e.g., by title, or author, or date, etc.) ranges from 1.00 to 2.00 cents per record. Merging of two or more information files to create a composite master file averages .70 cents per record. Unit costs for the MERGE program depend largely on the average length of each record. An alphabetically ordered vocabulary (thesaurus) listing of non-trivial words in any information field(s) averages 1.20 cents per record processed using the VOCAB program. An alphabetically ordered index of keywords and phrases, numerically indexed to all records in which each appears, averages 1.23 cents per record using the INDEX program.

File searching, which is the only high frequency continuing operation, averages .36 cents per record on file. Fluctuations in unit costs are comparatively minor and depend on both the complexity of the search formula and the number of information fields examined in the search.

Other costs associated with TRANSDAC center principally on the direct access storage of both programs and datasets. Under the existing billing algorithm, software storage charges are fixed at $2.62 per day. Online storage of the bibliographic dataset costs about $.88 per record per year.

Global cash costs attributable to system operations are roughly estimated at $7-8,000 at an annual rate, approximately 70 per cent of which originate with file searching activities ranging from zero to twenty job submissions each day.

System operations have been neither intensive nor extensive enough to suggest that the preceeding cost data are representative of all TRANSDAC applications. In particular, there is a large element of hidden cost associated with information monitoring and system updating which is implicitly allocated to personal time or economic research functions. Were the system to be established on a formal basis, dollar values would have to be assigned to these implicit costs, with a commensurate impact on total system costs.

TRANSDAC

SUMMARY

TRANSDAC is a computer based text storage and retrieval package providing a set of eight main program modules that can be selectively executed over remote terminals. It is characterized by;

(1) efficient use of IBM 360 hardware and its related Operating System: (2) direct access to disc-resident data files: (3) flexibility of open-ended file structures: (4) variable indexing, retrieval and vocabulary control features: (5) easy-to-use control statements: and (6) powerful search-profile options using Boolean and relational operators.

It has been established in the Canadian Transport Commission to effect the transition from manual to automated information management systems. Principal current use centers on bibliographic control of 1500 selected "high-interest" reference documents of continuing value to research programs. Total system content, as of April 1974, is 4500 detailed reference citations maintained on five separate information files.

REFERENCES

Yerke, T.B. et al. FAMULUS: A Personal Documentation System. Pacific Southwest Forest and Range Experiment Station, Box 245, Berkeley, California, 1969. 40p.