ROLE OF COMMERCIAL DATA BASES IN INFORMATION RETRIEVAL SYSTEMS — A CASE STUDY WITH CA SEARCH DATA BASE USING CAN/SDI PACKAGE

S KASIVISWANADHAM
INSDOC, Regional Centre
CSIR Complex
Taramani
Madras-600113

INTRODUCTION

This paper aims at discussing various aspects of commercial data bases being used in libraries and information centres for information retrieval services. The sources of these data bases and the tape services offered by multinational or national information systems have been dealt with. The various aspects of computer, such as, hardware and associated software and processing these data bases either off-line or on-line, have been described. In the batch mode process, the various operational aspects of SDI service rendered from INSDOC Regional Centre, Madras, from Chemical Abstracts tapes have been explained with flow chart.

The need for the creation of data base at institutional level such as, ENVIS (Environmental Information System), Madras Centre, where the size of the data is small and information needs are specific, has been emphasized.

COMMERCIAL DATA BASES, GROWTH

For the purpose of information storage and retrieval, a data base means a bibliographic data base in machine readable form usually available in tapes. The data contains bibliographical details (citations/abstracts or even full text of documents). The data stored in such data bases are identical to the hard copy version of various abstracting and indexing services.

The growth rate of these commercial data bases can be understood from the fact that DIALOG II in USA offers now six full text data bases, such as, Drug Information Full Text, and Harvard Business Review. It offers more than 50,000 password holders access to more than 200 data bases containing 100 million records on-line and available in over 80 countries. The citations are culled from more than 60,000 journals. Currently an estimated 2500 data bases are available.

Commercial Data Bases: Information Systems

The importance of commercial data bases in the computer-based information systems has been realised by major libraries/information centres. As a consequence, multinational information systems and national information systems have come into existence with the support of agencies like IAEA, UNESCO, WIPO, and UNIDO. These national and international information systems are mainly discipline-oriented, for example, INIS, AGRIS, and ISDS.

Information systems are also formed by groups of countries. Examples of such systems are: Space Documentation Centre (SDC) of European Space Agency (ESA). European Nuclear Documentation Service (ENDS) of European Atomic Energy Community (EURATOM).

Network information systems have come into existence making on-line access to remote data bases (i.e. data bases situated very far from the searcher/user of the system) feasible with the help of data communication networks.

The services from the data base tapes can be had either from the data base producers or vendors. However, services from data bases, such as, NTIS can be obtained only from vendors. Hence, service from vendors, such as, SDC (USA) is essential with reference to specific data bases.
Commercial Data Bases: Sources

The sources for commercial data bases are:

a. Information networks - multinational or national
b. Commercial Agencies
c. Others

The data bases produced by multinational or national information networks can not be purchased, leased or hired. The only way of availing the tape service from these data bases is to actively participate in the network system. Examples of these data bases are: 1) AGRIS 2) INIS and 3) MEDLARS.

Commercial Data Bases and Hardware Requirements

Commercial data bases can be processed with any main frame computer system such as, IBM 360/370. The hardware requirements may be 2 to 3 tape drives and one disk drive with on-line printer/terminal. The system should possess direct access devices such as, a) Magnetic Drum, b) Magnetic Card, c) Fixed Disk Store and d) Exchangeable Disk Store. The feasibility of processing the commercial data bases with the mini and microprocessor based computers may have to be studied. The experience of the Department of Library Science, Brigham University, for on-line search may be taken into consideration. According to Rajagopalan, “locally made microprocessor systems would practically solve the hardware requirements. There are at least 20 or more well-established vendors in the country at present. Some of the brands are already tried and tested for library applications. It is also claimed that they are compatible with the IBM-PC Series and the system may need Rs. 3 to 4 lakhs for installation”.[9].

Commercial Data Bases and Software

The operational aspects of any computer-based information system are mainly dependent on the availability of the existing software. The development of software is very costly and time consuming and efforts must be made to utilise the existing packages. Generally the associated software is available with the data base producers or vendors. The package has to be slightly modified depending on the computer system for installation. The package also needs modification when the formats of the tapes are changed. For example, the reformating programme of the CA Condensates tapes of CAN/SDI package installed at the Computer Centre, IIT, Madras, has been modified for converting C.A. Search data base into NRC MARC II format.

In the context of an institution providing information retrieval service with a single data base to limited users and having the availability of computer facility, it may be advisable to develop its own software. But, for national or international services, it is economical to use the software which deals with an environment suitable for multiple data bases such as, CAN/SDI. TEXTPAC, VEERA II, GETEL, ISIS, and STAIRS. With the promotional efforts of UNESCO, CDS/ISIS microversion and IV + V are becoming available now almost free of cost. MINISIS package can be run of HP 3000 System.

Commercial Data Bases: Processing

Off-line processing

These commercial data bases can be processed either off-line or on-line. In batch process, the user has to wait for his answer and the response time may be in the order of hours and days. Many organisations developed off-line information systems for their use and some organisations such as CISTI, Canada have developed software packages such as CAN/SDI. In India, the Indian Institute of Science, Bangalore, offers offline SDI service from INSPEC and BIOSIS.

The Indian National Scientific Documentation Centre (INSDOC), Madras Regional Centre, is providing SDI Service from Chemical Abstracts Search tapes using the IBM 370/155 computer located at IIT, Madras.

Operational details (Case Study)

As part of Current Awareness Service, Selective Dissemination of Information (SDI), from Chemical Abstracts Search Data Base tapes is being rendered from the INSDOC Regional Centre, CSIR Complex, Madras. The work flow of this batch mode operation is illus-
COMMERCIAL DATABASES

trated in the Flow Chart (Fig.1). Provision has also been made for generating the data base. The system uses CAN/SDI package and the computer jobs are run at the Computer Centre of IIT, Madras, on IBM 370/155. The system requires a core memory of 512k (Indian subset 386k) bytes. It can run under OS/MFT/MVT or OS/VS, OS/TSS compilers on IBM 360/370 computer systems. The peripheral configuration for the package is minimum three tape drives for 800/1600 bpi, a disc drive, a card-reader punch, and an on-line printer.

The entire operation is in five phases.

i) Profile generating/updating

ii) Data base conversion

iii) Searching data base with user profiles

iv) Sorting the searched answers

v) Supplying computer list of references at regular intervals.

Profiling

This has been represented in the first phase of the Flow Chart. The user profile is a list of terms representing the user’s interest in the language of the data base and a statement of the logical conditions required for a match. In a computer search, anything that is machine readable can be considered as a profile term and it consists of a word or a word fragment from a title or abstract, keyword, subject index term, journal coden, authors’ names, parts of the bibliography, and so on. The task of the profile construction is to choose those terms most likely to appear in the records of the data base.[3]

Construction of a profile is more a technique than an arrangement of keywords vertically and horizontally. It depends not only on the knowledge of the data base but also on the ability to define the subject matter in terms of the contents of the data base.

Profiling requires the following operations:

a. Analysing the topics

b. Gathering profile terms
c. Drawing concept maps
d. Coding profile terms
e. Formulating search expressions
f. Punching - preparation of profile decks
g. Compilation of profiles.

The compiled profile programme creates the user profile master file on a magnetic tape. This as well as the data base tapes are input to the search operation. (Sample profile listing - Fig 4)

Data Base Conversion

This has been dealt with in the 2nd phase of the Flow Chart. INSDOC Regional Centre, Madras, provides SDI service from the commercially obtained Chemical Abstracts Search Tapes.

CAN Search is a weekly computer readable service which is obtained from the American Chemical Society, Ohio. It provides references to chemistry and chemical engineering related documents abstracted in Chemical Abstracts and includes the control, bibliographic, and index data segments for all documents in all 80 sections. The odd numbered issues contain references on biochemistry and organic chemistry. The even numbered issues cover analytical, physical, macromolecular, and applied chemistry and chemical engineering. Over 15,000 journals, patents issued by 26 countries, new books, conference proceedings, and government reports are regularly monitored.

The tape format is in a variable length blocked form. The maximum record length is 3516 bytes, the maximum block size is 3520 bytes. Tapes are 9 track unlabelled. The CA Search tape is in SDF (Standard Distribution Format).

The conversion programme has been written in PL/I language. The programme first reads the SDF record and checks the copyright. The programme calls the external routine USASCII to translate ASCII characters into EBCDIC. The programme checks (CAN) citation abstracts number unlike (TAN) temporary abstracts number in C.A. Condensate tapes. The conversion programme for reformating CAC has been modified in order to reformat CA Search.

The controlling data element in the C.A. Condensate data base and C.A. Search Data Base are different.
START

Collect profile data

Punch profiles

Compile profiles

Profiles

next record till end file

SEARCH

Record hit

YES

NO

next key word combination (Logic)

next key word

Searched Answers

SORT

FEEDBACK ADDRESS

Sorted Answers

PRINT

OUTPUT

END

Fig. 1
### COMMERCIAL DATABASES

<table>
<thead>
<tr>
<th>Character position</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 to 4</td>
<td>Record length</td>
</tr>
<tr>
<td>5</td>
<td>Record status</td>
</tr>
<tr>
<td>6 to 9</td>
<td>Implementation codes</td>
</tr>
<tr>
<td>10</td>
<td>Indicator length</td>
</tr>
<tr>
<td>11</td>
<td>Identifier length</td>
</tr>
<tr>
<td>12 to 16</td>
<td>Base address of data</td>
</tr>
<tr>
<td>17 to 19</td>
<td>Length of 'length of data field' in each entry</td>
</tr>
<tr>
<td>20</td>
<td>Length of starting character position in each entry</td>
</tr>
<tr>
<td>21</td>
<td>For future use</td>
</tr>
<tr>
<td>22 &amp; 23</td>
<td>Number of characters 3</td>
</tr>
</tbody>
</table>

**Fig. 2**

**Field alternatives**
- 1)
  - Indicator length = 0
  - Identifier length = 0

**Fig. 2**
The record format for INSPEC file is based on the USA and British standards for communication of bibliographic information which in turn are based on the Library of Congress MARC format.

**Data Elements:**

<table>
<thead>
<tr>
<th>Sl. No.</th>
<th>Name of the data element</th>
<th>Length of fields</th>
<th>No. of bytes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Control No.</td>
<td>F</td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td>Record Type</td>
<td>F</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>Title of Record</td>
<td>V</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Subject Index Headings</td>
<td>V (with Variable subfields)</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Free Indexing Terms</td>
<td>V (with Variable Subfields)</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Treatment Codes</td>
<td>V (one to three bytes)</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Title of Corresponding Higher Level Publication</td>
<td>Optimal (V)</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Language</td>
<td>V</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Title of Conference</td>
<td>V</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Authors</td>
<td>V</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Editors</td>
<td>V</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Translators</td>
<td>V</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Abstract No.</td>
<td>F</td>
<td>8</td>
</tr>
<tr>
<td>14</td>
<td>Coden</td>
<td>F</td>
<td>6</td>
</tr>
<tr>
<td>15</td>
<td>Standard Book No.</td>
<td>F</td>
<td>10</td>
</tr>
<tr>
<td>16</td>
<td>Report</td>
<td>V (length free form)</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Patent</td>
<td>V (length free form)</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>Vol. &amp; Issue</td>
<td>V (free form)</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>Location of Conference</td>
<td>V</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>Place of Publication</td>
<td>V</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>Country of Patent</td>
<td>V</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>Pages of level 1 Type Record</td>
<td>V</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>Pages of level 2 Type Record</td>
<td>V</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>Inclusive Page</td>
<td>V (Free format)</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>Description of non Conventional Media</td>
<td>V (eq. § MAG. TAPE 400 FTH)</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>Author Affiliation</td>
<td>V (Free Text)</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>Editor Affiliation</td>
<td>V (Free Text)</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>Publisher</td>
<td>V (Free Format)</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>Organization Issuing Report</td>
<td>V (Free Format)</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>Inclusive dates of Conference</td>
<td>V (Free Format)</td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>Date of Publication</td>
<td>V (Free Format)</td>
<td></td>
</tr>
</tbody>
</table>
## COMMERCIAL DATABASES
### PROFILE LISTING

| T | A M *MEMBRANE | E 1 ARTIFICIAL | C 1 NATURAL | D 2 PHASE TRANSITION | F 1 FUSION | F 3 LIPID PROTEIN INTERACT* | G 1 LIPSONAL | H 1 LIPID VESICLES | J 2 UNILAMELLAR | K 1 LIPSONES | N 1 ULTRA STRUCT* | P 1 DYNAMIC | R 1 MODEL | S 2 TRANSP* | T 1 DNA | U 1 PIPIRON MICROSCCP* | W 1 DAMAGE | AA 1 ESTIMAT* | A 2 REPAIR | AC 1 LEHAL | AD 1 MUSTAGENIC EFFECT | AF 1 FURAZOLIDONE | AG 1 MUSTAGENISIS | AH 1 REASSOCIAT* | AJ 1 AGA ROSE CHROM* | AK 1 REPAIR MECHAN* | AI 1 PLASMID | AM 1 GENTIC* | AN 1 MOLECULAR | AR 1 ENDO | AT 1 EXONUCLEASE | AT 1 RADIATION BIOLOGY | AU 1 PHOTOBIOLOG* | AW 1 NEAR ULTRAVIOLET LIGHT | BA 1 GFI FILTRATION CHROM* | BB 1 ELECTROPHORESIS | BC 1 ULTRA CENTRIFUGATION | BD 1 VIBRIO | BE 1 CHOLORAE CELL | BF 1 RADIOBIOLOG* | BG 1 DYSUGEN |
| 1401 | 1401 | 1401 | 1401 | 1401 | 1401 | 1401 | 1401 | 1401 | 1401 | 1401 | 1401 | 1401 | 1401 | 1401 | 1401 | 1401 | 1401 | 1401 | 1401 | 1401 | 1401 | 1401 | 1401 | 1401 | 1401 | 1401 | 1401 | 1401 | 1401 | 1401 | 1401 | 1401 | 1401 | 1401 | 1401 | 1401 | 1401 | 1401 | 1401 | 1401 | 1401 | 1401 | 1401 | 1401 | 1401 |

**Fig. 4**
Searching

The third phase in the Flow Chart explains the search procedure. The users' queries in the form of search expressions search the data base and these search results are stored on to a magnetic tape. The search logic has been obtained through Boolean logic “AND” (+), “OR” (/), and “AND NOT” (-) symbols.

The converted data base tape and the profile tape are being used as input to the search programme.

The search programme reads the compiled profiles and matches them against the records on the NRC MARC data set. The programme is very sensitive to the format of the compiled profile data set and the format of the NRC MARC data set. The programme is coded in OS 360/F level assembler language for efficient operation.

The search will be made from all the access points in the data record, namely, author, title, classification code, type of the document, language, corporate author, coden, etc. The searched answer tape is input for sorting the searched answers.

Sort

The search programme gives all the hits on a file called SDI ANS. This file is sorted by passing on the required parameters to the IBM utility programme called SORT available in all OS.

Print

The print programme does the last job of the DI system, i.e., printing on hard copy. It puts each record in a predetermined format and the printout is in double punched card format. Each record is printed on both sides of the hard copy, i.e., left side and right side. The left side part is for users and the right side for feedback which consists one extra feedback questionnaire line “Is the citation useful” with options YES/NO/CANNOT TELL comment. Each reference is consecutively numbered. (Sample printout - Fig.5).

Advantages of the System

i. It minimises the time spent in locating and scanning the current literature,

ii. It extends coverage to publications not readily available,

iii. It produces highly personalised bibliographies needed by individual workers;

iv. It permits the expansion and narrowing of literature requirements, if any,

v. User can build his own personalised information file of references for current and future use;

vi. User is having literature from a range of access points like key words, title words, author's organisations and classification code.

Commercial Data Bases and on-line Processing

There are different types of on-line systems:

i. On-line enquiry systems

ii. On-line updating systems

iii. On-line time sharing systems and

iv. On-line interactive systems.

The user may have access to files from direct access devices, such as, magnetic disk and may obtain instantaneous response to his query.

In this system, access to remote data bases, i.e. data base situated very far from the searcher/user, is feasible because of suitable communication links between the terminal (used for searching) and the data bases being available. In fact, data communication networks such as, Tymnet and Telenet connecting various data bases and user terminals in USA and UK and some other countries have already been setup and are in use.

In Indian context, this can be performed by INDONET of Computer Maintenance Corporation and NICNET of the National Informatics Centre.

On-line Search - Indian Scene

In India, a project is in progress in National Aeronautical Laboratory (NAL), Bangalore.
Fig. 5
KASIVISWANADHAM

to set-up high-speed terminals known as MIKROTEL package with effective support of scientific user organisations in Bangalore like Defence Research and Development Organisation, CSIR, and UNESCO experimentally for 6 months envisaging the establishment of a 2400bps post and telegraph microwave link from Bangalore to Bombay and of an Overseas Communication Service and ITALCABLE satellite link through INTELSAT from Bombay to Frascati, Rome. DESIDOC (Defence Scientific Documentation Centre) has now a hookup with DIALOG costing about $70 per query.

COMMERCIAL DATA BASE(S) Vs DATA BASE GENERATION

The role of commercial data bases for information retrieval services is significant. But, generation of data base has become necessary in a library/information centre where the size of the data is small and specific, and the information needs of their users are specific in well defined areas. Environmental Information System, Madras Centre, can be cited as an example here. The centre is one of the first nine centres (later one more centre was added) selected by the Govt. of India, Department of Environment, for information retrieval services in the areas of: 1) Ecotoxicology, 2) Biodegradation of Wastes, 3) Environmental Impact Assessment, and 4) Environmental Systems Analysis. The centre is part of the (Centre for Environmental Studies), Anna University, Madras. The centre creates awareness and provides information services in the said areas.

To disseminate information in the aforesaid areas the Centre brings out a half yearly publication entitled Environmental Abstracts, and an annual publication entitled Environmental Data Base Survey which includes bibliographical details of primary sources available in its own collection as well as 12 other major libraries of Madras city and from its own collection in the above mentioned areas. It is also having a good report collection.

As part of current awareness service, the centre also brings out a publication entitled Current Literature Listing.

The Centre proposes to computerise their information system since the Centre is having access to an IBM 360 and a prime computer situated in Anna University, Madras. The access to a terminal is also feasible.

The important element in the proposed information retrieval system is to generate its own data base according to a standard format such as, MARC II. However, the format proposed by the International Standards Organisation is suited not only for bibliographic exchange but can also be used for data bases oriented to content-oriented search techniques (ISO 2709-1973). (Fig.2). The various data elements in a record are shown in Fig. 3.

The next important element in the proposed computer system would be the software in terms of the computer in use.

In this regard, it is worth considering the experiences of the Department of Library Science, Brigham University, for the proposed system.

According to Stirling, “A programme, Dialtwig, written in IBM PASCAL and operating on IBM PC, is used for many features of the DIALOG system. Structured search assignments are given to the students. These assignments are performed on an information science data base obtained from the School of Library Science (Berkeley) and on a Data-base (SLIS) reporting students research reports. In addition, other data bases are being created locally or down loaded from ERIC.

It also states that information storage and retrieval service can be rendered through IBM PC microcomputers. It states that originally 64k memory IBM PC be purchased with an expansion board. An additional 192k of memory can be added. It can have one disk drive 10 megabytes and two or three floppy drives. Special programming techniques, usually at the assembly language level, do permit efficient utilization of memory of 64k”[8].

If the proposed system uses an IBM 360, software packages such as, CAN/SDI and TEXTPAC can be used for processing the data base generated at the institution level. But the format of the data base should be in accordance with the package such as, NRC MARC II Or Textpac format as the case may be. In case of a prime computer, the package has to be altered or software has to be developed.

It is ideal for this Centre to have a microcomputer compatible to IBM PC. If it can
process the computer jobs through microcomputer, the software for information storage and retrieval service can be obtained from the manufacturers of the system or has to be developed.

CONCLUSION

The importance of commercial data bases in computer-based information retrieval systems for information retrieval services either from commercial data bases of from data bases generated at institutional level are preferred in view of world-wide information explosion, degree of specialisation in research and heavy increase in the price of the hard copy versions of the commercial data bases.

ACKNOWLEDGEMENT

The author thanks Shri T.S. Rajagopalan, Scientist-in-Charge, INSDOC, New Delhi, for his permission to present the paper. The author also acknowledges with thanks the Head, ENVIS, Madras, for his approval for incorporating his Centre's information dissemination activities. The author also acknowledges with thanks Prof. H.N. Mahabala, Computer Centre, IIT, Madras, Dr. M. Ramaiyah, Director, SERC, CSIR, Madras, and Shri P. Venkatesan, Central Library, IIT, Madras, for their comments.

REFERENCES: