DST SEMINAR ON

Application of Computers to Bibliographical Information Processing: Some Developments in India (Bangalore) (10-13 July 1978)

DATA BASE CREATION AND INTERNATIONAL FORMAT FOR THE EXCHANGE OF BIBLIOGRAPHICAL INFORMATION

R SATYANARAYANA, Scientist, INSDOC, New Delhi-110012

Discusses some of the definitions pertaining to the concept of data base. Outlines briefly the evolutionary pattern of data base concept. Considers some generic principles involved in data base design, delineating the problems involved in its creation. Explains the basic features of a bibliographical format. States that. standardisation in bibliographical format is essential for proper exploitation and exchange of information *in* the context of proliferation of machine-readable data bases in Science and Technology.

O INTRODUCTION

The role of an effective information system in accelerating the progress of Science and Technology is being increasingly realized in every country. Attempts are being made to develop computer-based information systems to meet the specific requirements. In India,

the DST has taken up the responsibility of developing a National Information System for Science and Technology (NISSAT). In this context, the 250 computers already available at different scientific and research institutions in the country are expected to be harnessed for this purpose. In the three tier system envisaged in the NISSAT plan, specialized information systems in different specialities such as leather, aeronautics, food technology, electronics etc. are envisaged with an integrated approach comprising facilities for exchange of information of all types with network pattern. This necessitates creation of data bases for individual subject areas internally and also use of commercially available data bases. An attempt is made in this paper to explain the data base concept as an important component of Information System, and some of the principles involved in its design and standardization.

I DATA BASE CONCEPT

II Historical Perspective

Historically the approach to data processing has been rather fragmented. The total data processing requirements for an organization have normally been split into a series of applications, with a separate file or files for each application. This practice has led to a proliferation of files, some with similar data. This not only created problems so far as updating and maintenance aspect is concerned, but also led to conflicting information being circulated due to delays in updating of

individual files. This situation led to the development of integrated approach in information processing. The data base concept is but a step in this direction. (8).

During recent years, data base has been given greater attention by computer manufacturers, users and academicians. In May 1969, a committee representing various groups held a series of meetings on data base and presented its recommendations in a report entitled "A Survey of Generalized Data Base Management Systems". This report is referred to as the CODASYL'S first report. The major objective outlined in this report was the examination of the feasibility of implementing a common (Generalized) data base system. Data definition, structure, levels of storage, management, and programming facilities were examined with regard to generalized data • base systems. In April 1971, the CODASYL Committee submitted a second report: entitled, "Data Base Task Group (DBTG)" which updated the earlier report. The DSTG report highlighted the Importance of implementing an independent dat base. (9) .

12 Approaches to Data Base Definition

The term data base has been defined in a number of ways in the literature of computer science. "Some apply the terms only to on-line processing, calling all of the files accessible to the on-line system the data base and, apparently leaving other files outside the base. Some

authors define it as a collection of files brought together as a single file commonly accessible by a given set of programs" (14).

The CODASYL Committee used the terms data base, data, bank, corporate data base, generalized data base, and common data base synonymously to indicate the pool of interrelated data pertaining to an entity, tailored to achieve efficiency in creating, updating, and retrieving data (7).

According to B.W. Romberg (20) "in its most basic form, a data base consists of a number of data elements, each of which is a unit of data that is complete in itself. These elements are organized into logically related groups called data structures. Information files, are composed of a large number of data structures of the same type."

The difference between data base systems and conventional files can be *seen* at this point. The data files in a data base system are organized in a fashion that permits their use in several applications, rather than a single application. Thus, in a data base system the focus shifts from a particular application and its specific input and output to a more general requirement for data files to serve a number of applications.

121 Bibliographical Data Base

Automated storage and retrieval has followed two main directions, bibliographic retrieval and data management. The data base concept discussed above was with reference to data management. Of course, the same could be used with advantage in bibliographical retrieval as well. In fact,research and development has gone into refining the techniques for indexing, abstracting, storing, retrieving and reporting of information from and about documents. As a result, large scale bibliographic data bases and retrieval services have been developed. Bibliographic data bases contain descriptive information about documents - books, periodical articles - the data typically included in these data bases is titles, authors, journal napes, volume and number, dates, keywords, abstracts etc. Inquiries about the data base are. intended to develop a list of document numbers or references which satisfy the needs of requester concerning a particular subject. The most common access to bibliographic data bases has been through large scale national services. Such data bases exist for numerous fields such as medicine, chemistry, biological sciences, engineering research and development. These services were initially batch-oriented. But, time sharing has enabled them to be offered on an interactive mode.

13 Data Base Design

As has been stated earlier, the data base concept is an advancement over the notion of an integrated set of files. "The design of d data base will provide: tables of descriptors and interrelationships (structure) separate from the data (content); mechanisms for accessing the data elements and data sets directly or indirectly via indices, pointers, conversion, or computation; a capability for movement of individual data items or groups from the associations in which they are stored to the associations in which they are processed, and back to the storage; protection of integrity and security of data items or groups of data items; management control of usage, content structure and optimization data that represents the activity of the enterprise,." (3).

131 Some General Principles

A necessary criterion in the creation of a data base - and one of its most essential characteristics - is that creators and users must agree on a common set of definitions for all the information in the data base, This is not so easy as it might seem. Elimination of data definitions in almost any medium to large-scale user of data processing immediately exposes three classic problems: 1) synonymous, or identical items of data called by different names in different application environments, 2) alternative definitions, or different systems using the same name to describe two different pieces of data; and 3) close definitions, or two different names

used to describe different pieces of data with definitions so similar that there should be only one name and one definition.

Therefore, as a first step, the data base designers should develop a complete dictionary of definitions. After they obtain the support of the users for these definitions they must keep it current and enforce it for new applications.

The next essential characteristic of significance in data base design and creation is deciding the data structure. Data structure has a bearing in the determination processing the data and access procedures. There are two fundamentally different approaches to accessing the data. These are sequential access organization and direct access. In sequential organization, records are located in sequence according to their key. This approach facilitates file maintenance but requires extensive searching to locate individual records. On the other hand, direct access permits the location of individual records without searching, but is much less efficient for maintaining a file with high activity. The storage media itself must be selected depending on the speed with which access to data is to be provided,

The security of information contained in the data base is of vital importance. This must be ensured by taking recourse to security hierarchy consisting of several levels of passwords.

Most of the above mentioned principles though intended to the creation of generalized data bases, could be used as guideline in the creation of bibliographic data bases.

14 Bibliographic Record Format

The bibliographical data bases such as MARC II, MEDLINE, INIS, CAC, INSPEC etc. store information about books, periodical articles, technical reports etc. The bibliographical description required to identify these types of documents is provided in varying degrees in each of these data bases.

The collection of information provided for a single document on machine-readable form as a self-contained unique logical structure is often referred to as bibliographic record format. The purpose of such a format is to identify the document unambigousiy and aid in its retrieval and also help the user in assessing its likely value in his search for pertinent information. Formal definitions for terms such as bibliographic description, bibliographic record, data elements, data fields, literature type and bibliographic level are provided in UNISIST Reference Manual(16) and are mostly accepted by the data base designers to a large extent.

141 MARC - II Format

As MARC - II format paved the way to the formulation of American National Standard for bibliographic information Interchange on Magnetic Tape (Copy on display).

this format is being used widely in indexing and abstracting services. • Therefore, the basic features of this format prove very helpful in designing a suitable format tailored to our requirements. As such, the salient features of MARC - II format are discussed here.

1411 MARC - II Is based on three elements: 1) The structure, 2) The content designators, 3) The content.

The structure provides the "basic machine framework to the record, while the content designators refer to the name by which the record could be identified. The content refers to the data recorded in the files.

1412 Components of Record Structure

There are three components - a) leader, b) record directory, c) variable fields.

Schematic Representation.

The leader provides particular information (such as the length of the record,-the type of record code or the material described in the record) about the ensuing record.

The Directory can be compared to the table of contents in a book. It indicates what the variable fields are in the record along with their location In the record. There is a 12 character record directory

for each variable field. The main elements present in the record directory are (a) tag, (b) length of field, (c) starting character position.

Record Directory

Tag is a three character code identifying the contents of a data field which corresponds to the directory entry.

Length of the field is a four digit number equal to the number of characters or bytes occupied by the data field which corresponds to the directory entry, including indicator and field separators.

The starting character position is a five digit decimal number giving the position of the first character of the data field which corresponds to the directory entry. The position is computed relative to the base address of the data part of the record. That is the start:: position of the first data field following the directory is zero.

Variable field: following the leader and directory, the record consists of variable fields; or data fields. Each variable field consists of alpha numeric data followed by a field separator.

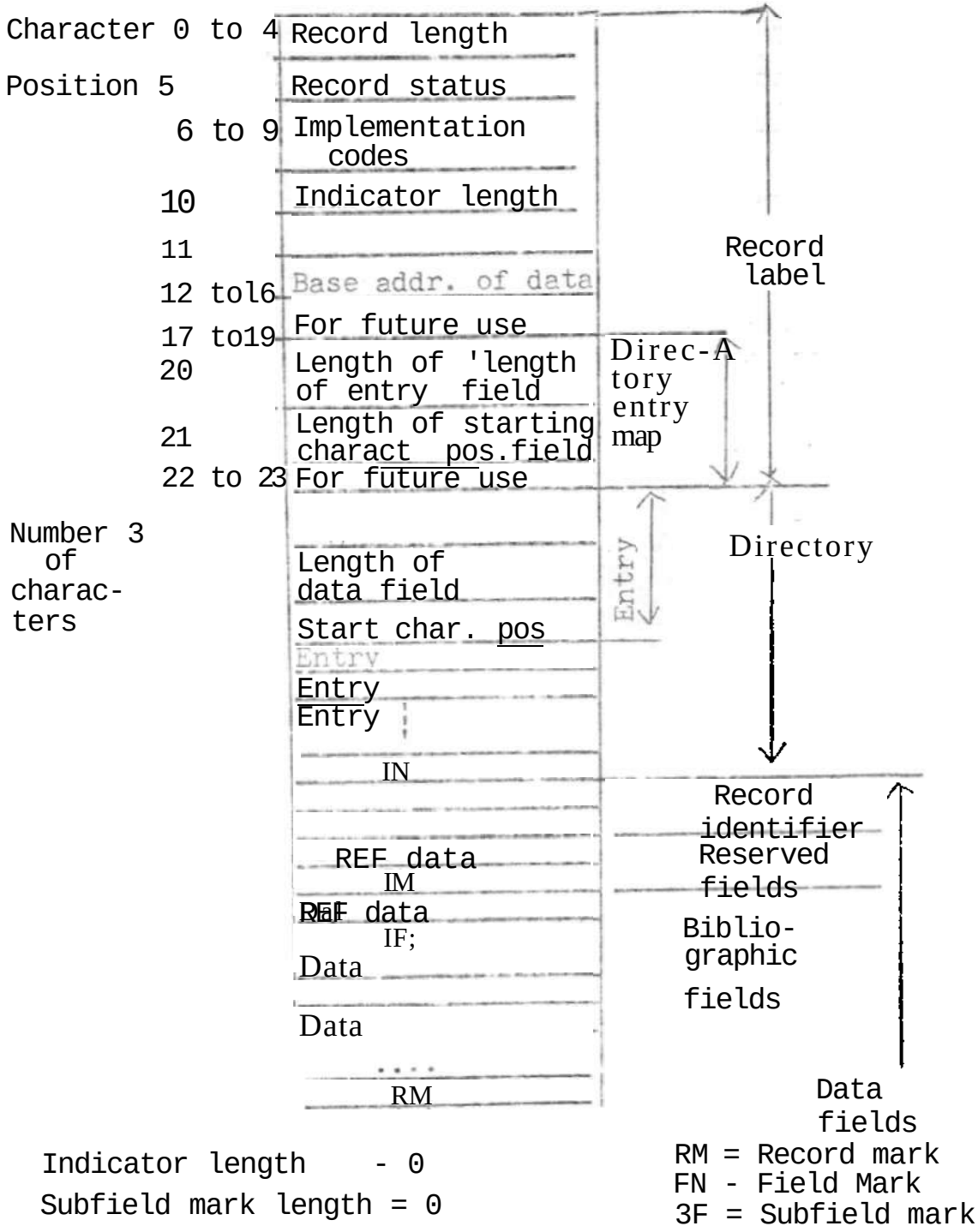
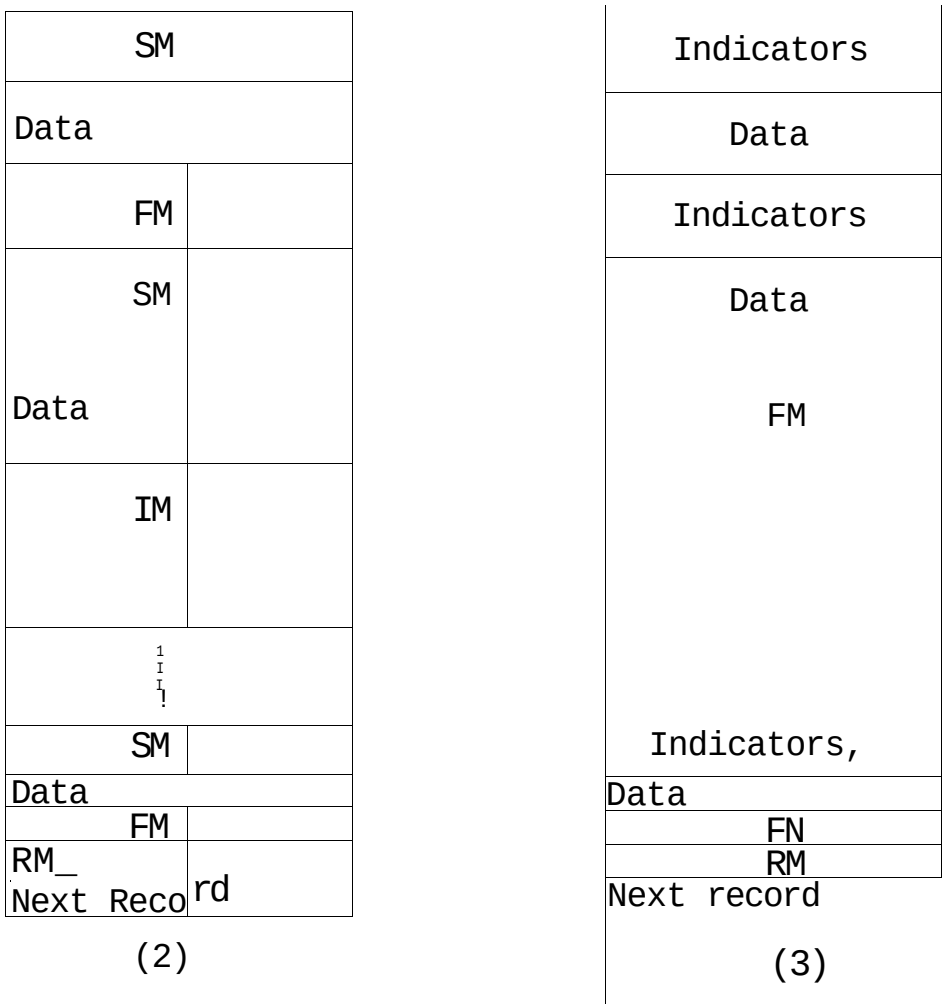


Fig: 1 - Representation of datn fields-in bibliographic record format. As per bS: 4748 : 1971.



Indicator length = 0 *0*

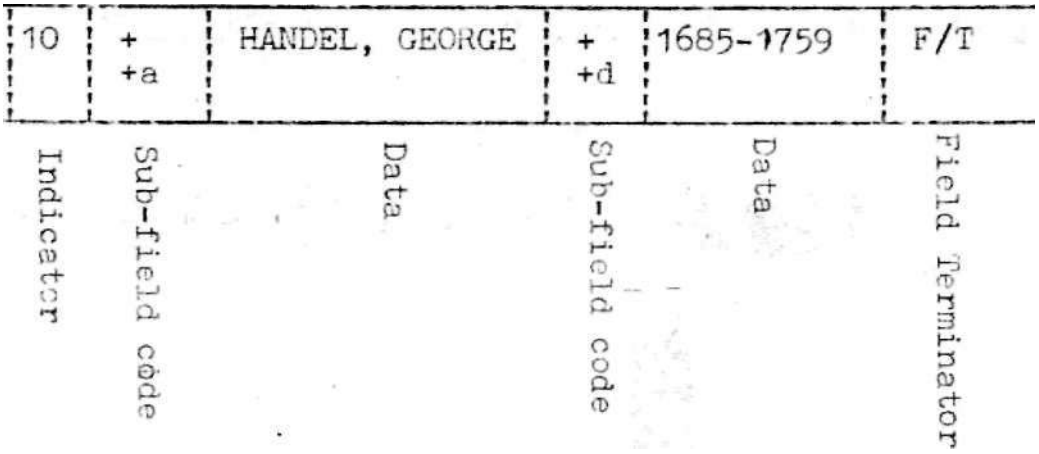
Subfield mark length 0 - 0

Fig: 1 - Representation of date fields in biblio-graphic record format. As per 35:4748:1971

RM = Record Mark
FM a Field Mark
SN = Subfield mark

International Format for Information Exchange

Graphic Representation of a Variable field



15 Need for Standardization

In any computer-based Information System, cost of producing data or information on machine-readable form constitutes major portion of the cost involved in project. It is only by sharing the data base among several user groups that COST effectiveness of the system could be achieved. This requires standardization in the design of a data base to reduce confusion and increase compatibility.

But, 'the question of establishing standards for use in the exchange of material between various centres is not the same as trying to agree to Universal System down to the last sub-field code. This is neither

possible nor desirable. Each specialized system will have its own characteristics and one must be careful not to overstep the mark .... Indeed a reasonable diversity should surely be encouraged .... It is enough to freeze the machine format which is the medium for exchange. We need urgently a common detailed implementation standard' (8).

The most important factor to standardize is the structure (content format) of a data base. "The structure of a machine-readable record may be likened to an empty container. Although, it is essential for transporting Its contents, it should impose few limitations on them. Ideally, the structure should be hospitable to all kinds of information, it should be independent of any particular hardware configuration, and it should be uniform regardless of the specific type of information it con tains"(19).

Formulation of standards for formats of fact files is somewhat difficult task compared to that of document or bibliographic data bases. The format proposed by ISC-2709-1973 is suited for (data bases designed to content-oriented search techniques. In this context tag oriented structure-is advocated for the data base design.

Some of national standard bodies such as British Standards Institution, American National Standards-Institute, ISO etc. have developed standards covering a) Bibliographic Information Interchange Format for Magnetic Tape (1), b) Magnetic Tape Labelling and File

structure for Information Interchange, and c) coded character sets for Information Interchange.

All these could be used with considerable advantage in the design of machine-readable data bases needed in the different tiers of NISSAT and at a later stage, when considerable practical experience has been gathered in India, the Indian Standards Institution may formulate its own standards to suit its national interest facilitating at the same time interchange of information on a global basis.

16 Conclusion

NISSAT envisages a network which facilitates* exchange of information of all types. Proper perspective of data base' concept is a basic requirement, if better results are to be achieved. We cannot minimise costs unless duplication of effort is avoided in the data base creation and software development.

In Indie, it is estimated that there are nearly 250 second and third generation computers installed mostly in scientific and research organizations, universities etc. These computer configurations are expec to be pressed into service for the purposes of developer ing NISSAT. In compatibility exists in card codes, character sots (6 bit or 8 bit) tape formats (7 or 9 track) and tape recording mode between the different generation manufactured by different companies. This requires standard data bases to facilitate information exchange both at national and international levels.

It is suggested that the experience which has been acquired at great cost, exists in different developed countries and it should be used to guide our national effort in this regard.

17 Acknowledgement

The author is grateful to Shri A. Krishnan, Scientist-in-Charge, INSDOC, for providing the necessary environment to prepare this paper.

18 Bibliographical References

1, AMERICAN NATIONAL STANDARDS INSTITUTE: American. National Standard for bibliographic information interchange on magnetic tape. ANSI 239.2-1971. New York, ANSI, 1971

2. AMERICAN "NATIONAL STANDARDS INSTITUTE: Magnetic rape labels for information interchange ANSI x 2-1969. New York, ANSI, 1969.

3 ADVANCES IN Information Systems Science, ed, by JT Tou. New York, PlenumPress, V.5- 1977 pp55-70.

4. BARETT (JW): Subject and mission oriented schemes: the international pattern as indicated by CAS, INSPEC, INI3 and others. Aslip Proc. 22 8; 1970; 386-94.

5. BRITISH STANDARDS INSTITUTE; 3S: 4748-1971: Specification for bibliographic information interchange format for magnetic tape.

6. CHOW (JV): What you need to know about DBMS.

(In. HOUSE (WC): Interactive decision oriented data base systems. New York, 1977. pp. 149-70).

7. CODASYL SYSTEMS COMMITTEE. Introduction to feature analysis of generalized data base management systems. GACM 14(5), 1971. pp.308-18.

S. COUGEft (JD) and McFADDEN (FR): Introduction to Computer-based information systems. New York. John Wiley, 1975. XVI,.654p.

9. COWARD (RD): MARC National and International Cooperation. 1973. The exchange of bibliographic data* and the MARC format. Verlag Dokumentation-Munche. P.17-26.

10. DAIA BASDTASK GROUP : "April 1971 Report" CODAS" Monroeville, Pennsylvania, 1971.

vi. DE GENNARO (R): A national bibliographic date base ir; machine readable form. Progress and prospects. Lib. Trends. 18, 4; 1970; 537-50.

12. ENGLES (RW) : Tutorial :on data base organization. IBM. Technical Memorandum, June 1969.

13. IRTERNATIOKAL STANDARDS ORGANIZATION: ISO/R 646-1973. 6 and 7 fait coded character set for information processing interchange.

14. ISO 2709: 1973. Documentation-format for bibliographic information interchange on magnetic tape.

15. LLEWELLYN (RW): Information systems. New Jeresy, Prentice-Hall Inc, 1976. xiv, 347 p. pp. 251-70.

16. MADEY (J), Ed.: Selected topics in information processing. IFIP-INFOPOL-76. Amsterdam North-Holland Pub Co., 1977. xvi} 534p. pp.449-79.

17. MARTIN (MD), Comp. UNISIST/ICSU-AB Working group. Reference manual for machine readable bibliographic descriptions. 1974. UNESCO, Paris.

',8. MARYAKA (LS) : Format recognition - a report at the LC. J Amer Soc Inf Sci 1971, 283-7.

'.9. MELTZER (HM) : Data base concepts and architecture for data base systems. 1969.

20. *RATHER (JC) : The realities of interchanging machine-readable bibliographic records. 1973. The exchange of bibliographic data and the MARC format Verlog Dokumentation Munchen-Pullech. P 27-36.

21. ROMBERG (BW): Data bases: there really is a better to manage your files. (In: HOUSE (WC): Interactive decision oriented drta base systems. New York, Petrocelli, 1977. pp.56-69.)

22. SCHOFFNER (RM): Some implications of automatic recognition of bibliographic elements. J Amer Soc Inf Sci 22,4; 1971; 275-82.

23. Standards Association Australia: Magnetic tape labelling and file structure for information interchange. A3 1068-1971.

24. YOUSSEF (L): Systems analysis **design.** Verginia, Reston pub co., 1975- pp. 106-21.