DRAWBACKS OF MARC FORMAT

RAVINDRA OJHA
Insdoc, New Delhi - 12.

MARC format in the course of years has become one of the most extensively used formats in the world for recording bibliographic data. The author points out some drawbacks of the format which he noticed while working with the INSPEC & CAC data bases.

INTRODUCTION

MARC format is currently being used extensively to store bibliographic information. Several data bases are already available in the market that use MARC format. Brief description of all recently published documents in a particular descipline are consolidated and stored in one magnetic tape. With the result there is considerable use of magnetic tapes for this purpose. As the scientific literature is doubling every ten years, we can safely assume that use of magnetic tapes for this purpose will also double at the interval of ten years provided the other parameters affecting the use of these tapes remain constant. Another significant observation is that several developing countries have started using the information supplied in these tapes to promote scientific developments in their regions. The number of records stored on tapes is growing and at the same time their use for different information activities is also increasing.

There are several versions of MARC format depending on the type of information being stored. Yet they don't differ significantly. Bibliographic information varies in length. Therefore, MARC format has presented a suitable way for representing bibliographic data. We shall try to examine the salient features of MARC format. The following are the basic structure of any MARC format.

1. Leader
2. Tag
3. Directory
4. Separaters, delimiters and terminators.

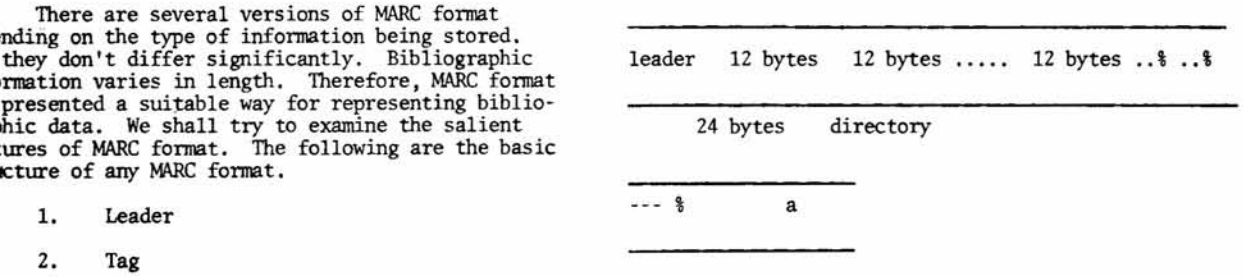
Leader: Leader is fixed in length. It is normally twenty four bytes long, and contains information such as total length of record, base address, type of data base, bibliographic level, etc. Base address indicates the point from where actual information starts.

Tag: To each type of data one tag is assigned, thus there is normally several tags in one particular record. Tags can be further subdivided by using delimiters depending upon requirement. Same tags may appear several times in the same record. Similarly in the same data base every record may not have all the tags.

Directory: Immediately after the leader, directory follows. For each tag there is at least one entry in a directory. There may be several entries in the directory for the same tag depending upon the occurence of the tag in a record. The length of one directory element is 12 bytes. The first 3 bytes indicate tag number, next four bytes indicate length of the information contained by this tag, and the last 5 bytes of the directory indicate relative position of this tag from the address contained in base address. These 12 bytes repeat in the directory for every tag present in the record.

Separater, delimiter, and record terminator:

These are special characters used in a record. Separaters are used to separate one tag entry from the others. Thus, when information under one tag is written then it is followed by a separator. Immediately after the directory there is one separater. Later on there is one separater for each directory entry. Delimiters are used to separate information contained in the same tag. They may be followed by one character to put various type of information in the same tag. Terminater are used to indicate end of the record. Thus for each record there is one terminator at the end. A MARC record will look something like as shown below.



While designing the MARC format care has been taken to cater to all types of requirements that may arise in describing bibliographic data. With the result record length and number of tags have grown excessively and some users are forced to make the data compact before using it. This gives rise to the necessity of conversion of data. Conversion may be done due to following reasons:

(a) To make the data suitable for a certain software.

(b) To change representation of data for particular hardware.

Normally, bibliographic data is too large and we are bogged down to conversion of one MARC format into another MARC format which means a large amount of C.P.U. time is consumed by conversion only. For

example, conversion of INSPEC tape to N.R.C. MARC format takes 60 minutes of C.P.U. time on IBM 370/155. This excessive wastage of computer time is basically due to the drawback of MARC format which is not amenable to different software.

While comparing the terms of each MARC record we ignore the presence of the same term in any other record. With the result a term which is present in different MARC records is compared with the same term in profile several times which results in unnecessary comparison of same terms again and again.

Let us assume we are searching 100 profiles each with 30 words on an average. Thus, we are searching nearly 3000 words. In this 3000 words some may be duplicate and they need be searched only once. For simplicity we assume that there is no duplicate term in the profiles. On the other hand two issues of Chemical Abstracts contain 15,000 records. If each MARC record contains 15 key words which can be compared with words in profile then the total number of words available for comparison is 225,000. This is excessively large and chances of duplicate words in MARC record is much more which is evidenced by the fact that in a profile, the same group of words select several records.

Secondly, by removing indentical words in a MARC record we can save in storage which will reflect in increased search efficiency.

The above data is far more conservative, because, normally there are much more words which can be compared hence this way of representing data is inefficient.

MARC format is dependent on directory, delimiters and separator which do not carry any information except that they lead us to a particular type of data. The average total length of MARC record and percentage of space occupied by separators, delimiters, terminator, directory and leader are given below:

Average record length = 609 bytes

Average area occupied by delimiters, separators, terminators, leader, directory = 228 bytes
Percentage of area occupied by leader, directory, separators, etc. = 37.438%.

It is really surprising to know that in any 100 bytes of MARC record 37.438 bytes are occupied by characters which don't carry any information. If we take two issues of C.A. Condensates with an average record length of 609 bytes and about 15000 records then total space occupied by the directory and other terms of this type is 3.6 million bytes. This large amount of storage will definitely result in longer search time and more storage.

Normally, we represent two types of data i.e. fixed length and variable length. Fixed length data are generally ISBN, ISSN, Vol. issue and date of publication. They can be represented without any directory and tag.

Preparation of profiles for querying the data base for SDI service is such that we use mostly keywords and titles for searching a data base as can be seen from the following:

* Term type	A	.70%	(Personal Author)
Term type	B	.10%	(Corporate Author)
Term type	C	.40%	(Journal)
Term type	K	3.5%	(Article, language, subject section, etc)
Term type	T	95.30%	(Title and Keyword)

It is evident that most of the time we are searching term type T which necessiates that data represented by T should be made readily available to satisfy 95% of query. MARC format treats all types of data equally.

CONCLUSION:

We have seen some drawbacks involved in the MARC format. Obviously this is not a format which is going to last long. Any universal format for bibliographic data base should be organised in such a way so that minimum conversion of data is required. When we move from one system to another system data should be organised in such a manner that if any user wants to examine any particular aspect he should not be bothered about other details. MARC format in its present state has tried to answer all types of user queries with the result that it has become unmanageable on a modest computer.

If we want to search a group of issues together then it is almost impossible to manage such a large amount of data.

ACKNOWLEDGEMENTS:

The author is highly obliged to SIC, INSDOC for having permitted him to publish this work. I am also grateful to my colleagues, who encouraged me for this work.

* These data were obtained from corresponding tapes after developing programs to obtain and analyse them.