CONTENT DESIGNATORS FOR MACHINE-READABLE RECORDS: A WORKING PAPER

Henriette D. AVRAM and Kay D. GUILES: MARC Development Office, Library of Congress, Washington, D.C.

Under the auspices of the International Federation of Library Association's Committees on Cataloging and Mechanization, an International Working Group on Content Designators was formed to attempt to resolve the differences in the content designators assigned by national agencies to their machine-readable bibliographic records. The members of the IFLA Working Group are: Henriette D. Avram, Chairman, MARC Development Office, Library of Congress; Kay D. Guiles, Secretary, MARC Development Office, Library of Congress; Edwin Buchinski, Research and Planning Branch, National Library of Canada; Marc Chauveinc, Bibliothèque Interuniversitaire de Grenoble, Section Science, Domaine Universitaire, France; Richard Coward, British Library Planning Secretariat, Department of Education & Science, United Kingdom; R. Erezepky, Deutsche Bibliothek, German Federal Republic; J. Poncet, Bibliothèque Nationale, Paris, France; Mogens Weitemeyer, Det Kongelige Bibliotek, Denmark.

All working papers emanating from the IFLA Working Group will be submitted to the International Standards Organization Technical Committee 46, Subcommittee 4, Working Group on Content Designators.

Prior to any attempt to standardize the content designators for the international exchange of bibliographic data in machine-readable form, it is necessary to agree on certain basic points from which all future work will be derived. This first working paper is a statement of: 1) the obstacles that presently exist which prevent the effective international interchange of bibliographic data in machine-readable form; 2) the scope of concern for the IFLA Working Group; and 3) the definition of terms included in the broader term "content designators."

If an international standard format can be derived, it would greatly facilitate the use in this country of machine-readable bibliographic records issued by other national agencies. It should also contribute significantly to the expansion of MARC to other languages by the Library of Congress. At *present, the assignment of content designators of most national systems is so varied that tailor-made programs must be written to translate each agency's records into the United States MARC format. The international communications format might become the common denominator between all countries, each national system maintaining its own national version.*

INTRODUCTION

The International Organization for Standardization standard for bibliographic information interchange on magnetic tape (1) has recently been adopted, following on the adoption of the American National Standard (2). These events, along with the implementation of the United States and the United Kingdom MARC projects and similar projects in other countries, have emphasized the importance of the international exchange of bibliographic data in machine-readable form.

There are many problems to be resolved before we can approach a truly universal bibliographic system. Many of these have been described in an article by Dr. Franz Kaltwasser (3). Basic to the exchange of bibliographic data is the requirement for an interchange format which can be used to transmit records representing the bibliographic descriptions of different forms of material (such as records for books, serials, and films) and related records (such as authority records for authors and for subject terms).

A format for machine-readable bibliographic records is composed of the following three elements:

1. The structure of the record, which is the physical representation of the information on the machine-readable medium.
2. The content designators (tags, indicators, and data element identifiers (4)) for the record, which are means of identifying data elements or providing additional information about a data element.
3. The content of the record, which is the data itself, i.e., the author's name, title, etc.

OBSTACLES

The structure of the record, as described in ANSI Z39.2-1971 and in the ISO standard on bibliographic information interchange on magnetic tape, has been fairly well accepted by the international bibliographic community. However, events have shown that as the different agencies examine their requirements and establish the content of their machine-readable records, the content and the content designators so established are not the same across all systems. This lack of uniformity is the result of at least four principal factors:

1. The different functions performed by various bibliographic agencies.

Bibliographic services are provided by many types of organizations issuing a variety of products. These products are dissimilar because the uses made of them vary, reflecting dissimilarities in the principal functions of the agencies involved. The main products of some of the different bibliographic services are briefly described as follows:

Catalogs serve to index the collections of individual libraries by author, title, subject, and series. To enable a user to find a physical volume rather than merely a bibliographic reference, catalogs also provide a location code. A unique form of entry for each name or topical heading used as an access point is maintained by means of authority files. The various access points serve to bring together works by the same author, works with the same title, works on the same subject, and works within the same series. A unique bibliographic description of each item makes it possible to distinguish between different works with the same title, and different editions of the same work.

National bibliographies provide an awareness service for those items published within a country during a given time period. A national bibliography is not a catalog, since it is not based on or limited to any single collection, nor is it concerned with providing access to the physical item itself.

Abstracting and indexing services are principally concerned with indexing technical report literature and individual articles from journals and composite works. Because these services generally index more specialized materials and are aimed at the specialist in a particular discipline, more complete indexing by means of a relatively large number of very specific subject terms is the rule. Like the national bibliography, the abstracting and indexing service is not concerned with a single collection or, in most cases, with providing access to the item on the shelf.

2. The lack of internationally accepted cataloging practices.

The Paris Conference of 1961, which resulted in the Paris Principles, set the framework for an international cataloging code. Following the conference, progress in standardization was evident in the work begun on the formulation of cataloging codes embodying, in varying degrees, the Paris Principles. One such code is the Anglo-American Cataloging Rules (AACR) (5). However, we are concerned with the present, and the differences that exist in the cataloging codes of various countries do create differences in the content that may affect content designation of machine-readable bibliographic records.

The differences between cataloging rules practiced in the library community and in the information community (6) are even more prominent. In the United States, these differences are clearly seen in a comparison between AACR and the COSATI rules (7). Even more significant is the fact that in preparing entries for abstracting and indexing services, it is common practice to use a name as it appears on the document, without attempting to distinguish it from names of other persons so as to bring together the works of a single author. In addition, cataloging practice in the information community often requires inclusion of data elements that are not used in the library community (e.g., organizational affiliation). It is obvious that these differences in practice are serious obstacles to achieving agreement on details of content designation for machine-readable records used in each environment.

3. Lack of agreement on organization of data content in machine-readable records in different bibliographic communities.

Bibliographic data can be organized in machine-readable form in many different ways. For example, one approach could be the grouping of data elements by bibliographic function, such as main entry, title, etc.; another approach could be the grouping together of information by type, such as all personal names, all corporate names, etc. There are pros and cons associated with each of these groupings. This difference in organization exists in some instances between the library community and the information community. For the present discussion, it is not appropriate to analyze the relative merits between the two points of view. It must be emphasized, however, that there is no optimum organization, and that a variety of users will use the data in a variety of ways. It is certainly true that any given system can define, upon agreement of its members, a particular use to be made of the data exchanged and, in this case, perhaps an optimum data organization can be defined ("perhaps" is used because hardware is another variable that comes into play).

4. Lack of agreement as to the functions of content designators.

There is a lack of agreement as to the functions of content designators, as well as a misunderstanding, in some instances, of the rationale for the assignment of certain of them to specific data elements. The lack of agreement as to the functions of content designators is clearly seen when one examines the use of the data element identifiers in the different national formats.

For example, in some cases the data element identifier is assigned to the data element according to its value in a collation sequence (e.g., a is smaller than b, b is smaller than c). The result is a prescribed order, from the smallest value to the largest, for selecting the data elements to build a sort key for file arrangement. In other systems, the data element identifier assigned to a data element is for the unique identification of that data element. There is no prescribed ordering built into the data element identifiers; the identification of the data elements allows them to be selected according to the requirements of the user to build a sort key for file arrangement. Data element identifiers in some cases are tag dependent, i.e., they identify the same data elements consistently when used with a particular tag and data field, regardless of the combination of data elements present in the data field for any particular record. In other cases, the data element identifiers are tag, indicator, and data dependent, i.e., the meaning of the data element identifiers changes and the data element identifiers are assigned to different data elements, depending upon the combination of data elements occurring in a data field for a particular record.

SCOPE

The scope of responsibility for the IFLA Working Group is to investigate the present assignment of content designators for the purpose of determining those areas in which there is uniformity of assignment and those areas in which there is not uniformity. Once this has been done, the Working Group's next task is to explore how best these differences can be accommodated so as to arrive at a standard for the international interchange of bibliographic data. Within that scope, the Working Group will first be concerned with the requirements for the international library community, i.e., libraries and national bibliographies. The magnitude of this assignment is such that it appears unwise to impose the additional problems of the needs of the information community concurrently. If the attempt is made to do so, and the result of the effort is failure, it will not be clear whether we failed because the task was too difficult or whether it is not possible to merge two communities with significant variation throughout their systems. On the other hand, if only the library community is approached at this time, the result of the effort can be success; but if the result is failure, at least one factor will be clear if only in a negative sense: there will be no lingering question as to whether the attempt might have succeeded had the problems of only one community been addressed at one time.

In summary, it may be stated that our attempt to standardize content designators within the library community will be complicated by: 1) the lack of an international cataloging code; 2) the dissimilarities in the products of various agencies created by the different functions performed by those agencies; and 3) the lack of an agreement on the functions of the content designators themselves. The lack of agreement on an international cataloging code will have an impact on our work, but is an area which is out of scope for the Working Group, and therefore can be considered a variable over which there is no control. The dissimilarities in the functions of the different bibliographic services are also a given. However, since it was possible to work around these differences in the formulation of the International Standard Bibliographic Description, it may be possible to do so for the standardization of content designators. Therefore, within the two variables given above, our emphasis should be placed on attempting to resolve the lack of agreement on the functions of content designators and then we can proceed to attempt to standardize the assignment of tags, indicators, and data element identifiers.

The present paper concentrates on the substance of the problem, namely, a statement of the definition of tags, indicators and data element identifiers and their functions, i.e., the information they are intended to provide to a system processing bibliographic data.

The concept of a SUPERMARC has been discussed in the literature (8, 9) as an international system for exchange, leaving the various national systems as they now exist. Each country would have an agency that would translate its own machine-readable record into that of the SUPERMARC system; likewise, each agency would translate the SUPERMARC record from national bibliographic systems into its own format for processing within the country concerned. At the international level, there would be only one record format. This concept has the theoretical advantage of eliminating the difficulties inherent in seeking agreement internationally. However, what has not been addressed is the problem inherent in this concept, namely, the problem associated with any switching language. This may be illustrated in the following manner. Consider the case of a national agency (called System 1) whose format is not detailed in regard to content and/or content designation. When System 1 translates to SUPERMARC, the result will be a SUPERMARC record, but it will be a SUPERMARC record still only defined at the level of detail of the limited record of System 1. This will be true regardless of the level of detail at which SUPERMARC is originally defined. Likewise, when a national agency (called System 2) accepts records from System 1 via SUPERMARC and translates the SUPERMARC records into its own format, the resulting records will be the limited records of System 1, regardless of the detail of System 2's local records. This may be schematically represented as follows:

System 1 (little detail)	SUPERMARC (great detail)	=No more detail than System 1
SUPERMARC (record from System 1)	System 2 (great detail)	=No more detail than SUPERMARC record from System 1

The result of this analysis suggests that systems with formats of less detail than that of SUPERMARC must permanently upgrade their national formats to the level of detail of SUPERMARC while systems with formats more detailed than SUPERMARC must be prepared to accept the fact that records from other countries will probably require significant modification. Therefore, although national variation is allowed in a SUPERMARC system, the international community still faces all the problems of international agreement, i.e., arriving at an acceptable level of content designation for SUPERMARC.

CONTENT DESIGNATORS

Bibliographic records in machine-readable form permit the manipulation of data and allow greater flexibility for the creation of a variety of products. The full potential of machine-readable files has not been exploited to date, but based on experience and the projection of this experience into the future, it may be said that the variety of uses of machine-readable cataloging data will be limited only by the imagination of the user. Among the possible products are printed catalog cards, book catalogs, special bibliographies, special indexes, book preparation materials, CRT display of cataloging information, management statistics (analysis of data by type of material, subject, language, date, or other parameters), etc. All of the above are possible in a wide variety of output formats.

In order to produce these various tools, there are four basic operations (10) which are performed on the data.

1. Store—the storage operation is the internal (to the computer) management of the data, i.e., how files are organized, the type of accessing technique(s) used, and the data elements (e.g., author, title) selected as keys to the complete bibliographic record.
2. Retrieve—the retrieval operation is used here in its broadest sense, to cover the following kinds of retrieval: the retrieval of a single element from a record; the retrieval of a known item, such as the selection of a record by unique number or author and title; the retrieval of a category of records, such as those for all French language monographs on a particular subject with an imprint date of 1968 or later; the retrieval of all bibliographic records for a particular form of material, e.g., serials. (The latter retrieval capability allows segmentation of files not only for display purposes but also for the implementation of certain file organization techniques.)
3. Arrange—the arrange operation puts information in a sequence that is most useful for the user of the product, i.e., an alphabetic sequence or a systematic arrangement.
4. Display—the display operation as used in this context implies formatting, the purpose of the operation being to make the information human-readable, e.g., display on a CRT, computer printout, and photocomposed output.

For example, to display a particular catalog record on a CRT device, the record must be retrieved from the data base by a known number or other means of access and formatted for display; or, to prepare a special bibliography, all records satisfying a particular search argument are retrieved from the data base, arranged in some predefined order, formatted and printed. The storage operation is implicit in the examples.

In order to perform these four basic operations through machine manipulation, content designators are assigned to the data content of the record. Therefore, it may be stated that the function of content designators is to provide the means for the user to store, retrieve, arrange, and display information in a variety of ways to suit his needs.

There are three types of content designators currently in use: tags, indicators, and data element identifiers. For the purposes of standardization, agreement must not only be reached on the definition of those three elements but also on other basic issues. The definitions for the elements are given below, as well as a general discussion of some of the decisions that must be made concerning each of the elements, prior to attempting to achieve standardization.

1. A tag is a series of characters used to identify or name the main content of an associated data field (11). The designation of main

content does not require that a data field contain all possible data elements (units of information) all the time. For example, the imprint may be defined as a data field containing the data elements, place, publisher, date of publication, printer, address of printer. The tag for the data field called imprint would be the same if only a partial set of the data elements existed for any single occurrence of the data field in a bibliographic record. Should the method of assigning tags be simply to assign a unique series of characters to a data field whereby the characters have no meaning other than to name the main content of the data field? Or is it desirable to give values to the characters making up the tag? In the latter case, a tag may identify a data field both by function and type of entry, thus allowing greater flexibility in internal organization of the data as well as its formatting for output.

2. An indicator is a character associated with a tag to supply additional information about the data field or parameters for the processing of the data field. Indicators are tag dependent because they provide both descriptive and processing information about a data field. Should alphabetic characters as well as numeric characters be assigned to indicators? Should the character ƀ (blank) always mean a null condition and the character ∅ (zero) have a value or a meaning? Should indicators with the same values and meanings be used for different data fields and their associated tags where the situation warrants this equality? For example, a personal name may be a main entry, an added entry, or a subject entry. If it is deemed desirable to further describe the type of personal name such as forename, single surname, multiple surname, or name of family, the indicators set for each of the data fields mentioned above would have the same value and the same meaning. This technique has the advantage of simplifying machine coding for the processing of different functional fields containing the same types of entries.
3. A data element identifier is a code consisting of one or more characters used to identify individual data elements within a data field. The data element identifier precedes the data element which it identifies (12). Should data element identifiers be given a value, i.e., file arrangement value, other than the identification of the data element? Should data element identifiers be tag dependent only or tag, indicator, and data dependent? Should the same data element identifiers be assigned, so far as is possible, to the same data element regardless of the field in which the data element occurs? Should data element identifiers be restricted to alphabetic characters or should they be expanded to allow the use of numerics and symbols?

The assignment of a filing value to a data element identifier is intended to minimize the effort required to create software for filing. However, assigning filing values to data element identifiers results in identifiers that are tag, indicator, and data dependent. On the other hand, without assigning filing values to the data element identifiers and using computer filing algorithms, the system can avoid data dependent codes, thus ensuring maximum consistency across all fields. For example, the use of the same data element identifier assigned to a title wherever a title appears in the record allows the flexibility of selecting all titles by data element identifier. Furthermore, tag, indicator, and data dependent data element identifiers create additional complexity in the editing procedure (13).

Although fixed fields are not content designators, they do take on similar characteristics as to function, i.e., to provide the means for the user to store, retrieve, arrange, and display information in a variety of ways to suit his needs. Therefore, they should be considered by the Working Group along with the content designators. A fixed field is one in which every occurrence of the field has a length of the same fixed value regardless of changes in the contents of the field from occurrence to occurrence. The contents of the fixed field can actually be data content, e.g., date of imprint; or a code representing data content, e.g., type of illustration; or a code representing information about the record, e.g., language of the record; or data concerned with the processing of the record, e.g., date entered on file.

Here again, certain basic issues must be resolved. Should the character b̸ (blank) be used to signify a null condition, e.g., in a record without any type of illustration b̸ (blank) would be used? Should the codes that represent more than two possible conditions be alphabetic or numeric? Should the characters 1 (one) and Ø (zero) be used to indicate an on-off condition, e.g., a book contains an index to its own contents (1) or it does not (Ø)?

It is important to keep in mind the eventual necessity of correlating the content designators and fixed fields for all the formats defined for different forms of material (books, serials, maps, films, music, etc.). By adhering as much as possible to the same content designators and fixed fields, the processing of different forms of material will be facilitated in terms of the software required to perform a particular process and to combine all forms of material in a single product, such as a book catalog.

REFERENCES

1. International Organization for Standardization. *Bibliographic Information Interchange—Format for Magnetic Tape Recording.* Draft international standard ISO/DIS 2709. Technical Committee ISO/TC 46 Secretariat (Germany), 1972.
2. American National Standards Institute. *American National Standard for Bibliographic Information Interchange on Magnetic Tape.* ANSI Z39.2-1971. New York: American National Standards Institute, 1971.
3. Franz Georg Kaltwasser, "The Quest for Universal Bibliographical Control," *Unesco Bulletin for Libraries,* 25:252-259 (Sept./Oct. 1971).
4. Data element identifiers have more commonly been referred to as subfield codes.

5. *Anglo-American Cataloging Rules.* Prepared by the American Library Association . . . North American Text. Chicago: American Library Association, 1967.
6. The term bibliographic services has been used to include all agencies concerned with bibliographic products. For this paper such agencies have been further subdivided into two communities: the library community, defined as including libraries and national bibliographies; and the information community, defined as including the abstracting and indexing services. This broad definition has been used for the sake of simplicity.
7. Committee on Scientific and Technical Information. *Standard for Descriptive Cataloging of Government Scientific and Technical Reports.* Washington: Committee on Scientific and Technical Information, 1966.
8. R. E. Coward, "MARC: National and International Cooperation," *in* International Seminar on the MARC Format and the Exchange of Bibliographic Data in Machine-Readable Form, Berlin, 1971: *The Exchange of Bibliographic Data and the MARC Format,* (München-Pullach, 1972), 17-23.
9. Roderick M. Duchesne, "MARC and SUPERMARC," *in* International Seminar on the MARC Format . . ., p. 37-56.
10. These basic operations are not used in this context to mean basic machine operations such as add, subtract, multiply, and divide.
11. A data field is a variable length field containing bibliographic or other data not intended to supply parameters to the processing of the bibliographic record, i.e., content data only.
12. There are in existence formats in which the data element identifier is a single character, i.e., a delimiter. In this case, there is no explicit identification function built into the data element identifier. If, in the particular data field, the data elements are all of the same type, such as a multiname data field, then the meaning of the delimiter is implicit.
13. Editing is used in this context to mean the human or machine assignment of content designators.