BIBLIOGRAPHIC/MARC* PROCESSING SYSTEM

System Manual

Compiled by Victor J. Schoenbach

James G. McClure, Project Director

Bates Buckner, Director
Technical Information Service

Thomas L. Hall, Director
Carolina Population Center

University of North Carolina at Chapel Hill

April 1977

* MARC is the Library of Congress acronym for Machine Readable Catalog.

Inquiries should be addressed to:

Cheryl Ward, Systems Librarian
~~James G. McClure~~
~~Technical Information Service~~
Carolina Population Center
University of North Carolina at Chapel Hill
University Square 300 A
Chapel Hill, North Carolina 27514
U.S.A. (919) 933-3006

also: Martin Dillon, PhD
School of Library Science

Victor J. Schoenbach
Department of Epidemiology
933-1258

Table of Contents

Introduction

The Bibliographic/MARC Processing System was developed to produce and maintain the machine-readable catalog of the Technical Information Service Library of the Carolina Population Center. During the four years the system has been in operation, there have been substantial expansions of capability, efficiency, and ease of use. The system's flexibility has enabled it to be used not only to produce title and subject catalogs of the TIS-CPC Library holdings as well as many special catalog-derived products, but also to maintain the TIS-CPC Population/Family Planning Thesaurus, bibliographies for special (non-TIS/CPC) collections, manpower training records for the United States Agency for International Development, and the catalog of machine-readable data files of public opinion surveys held by the Social Science Data Library of the Institute for Research in the Social Sciences.

File structure

At the core of the system is a master file, whose structure is modeled on the MARC (MAchine Readable Catalog) format developed by the Library of Congress. The essential characteristic of MARC-format records is that they can accomodate a varying number of variable length data items, affording considerable generality of use.

Each MARC record begins with a 92 byte section, which contains the record id number (called the "did" -- document identification), the creation date, and various other fixed length data and control items including a count of the number of variable length data fields that follow.

A record directory follows the fixed field section. Entries in the directory correspond one for one with the variable fields they point to and appear in the same order. Each entry contains the address of the beginning of the variable field, its length, and information about the type of data it contains. Additional descriptive information is placed in the first bytes of the field itself.

Each field is associated with a three-letter code called a tag. For example, in the TIS-CPC Library catalog, "til" refers to the document title, "cal" to its call number, "imp" to its imprint. In the USAID Manpower Training Record file, "nme" is the individual participant's name, "nat" his or her nationality, and "spn" the sponsor of the training. Though tags are generally defined when an application is being structured for use of the Bibliographic/MARC Processing System, new tags may be defined any time a need for them is perceived.

Any individual record may contain from one to one hundred tagged fields. A record may also contain more than one field of the same type. For example, a document might be associated with several primary subject terms ("sup"), several primary geographical area references ("gap"), etc. Multiple occurrences of the same tag in a record are differentiated by tag sequence or site numbers.

Each variable field may be further subdivided into variable length subfields. Each subfield is associated with a letter of the alphabet (the subfield code). A field may consist of a single subfield or as many as 12 subfields. A

TIS-CPC Library example would be the imprint (tag "imp"), where the place of publication is subfield "a", the publisher subfield "b", and the year of publication subfield "c".

Finally, tagged fields of the same type may be further distinguished by the use of a tag suffix or indicator. Indicators serve to denote some minor difference between two fields that in other respects should be treated identically and which have therefore been assigned the same tag. A TIS-CPC Library example is the use of the indicator "a" for a title whose first word (a non-English language article) is to be disregarded in sorting and filing (English-language leading articles "A", "An", and "The" are automatically disregarded). In other respects, the English and other language titles are treated identically.

Processing capabilities

Processing capability of the system lies primarily in the areas of information storage, retrieval, and display. The MARC-format record structure can accomodate most types of cataloging data as well as any textual information that can be structured into non-hierarchical records (some limited processing of hierarchical records is available, but mostly through special program routines written expressly for the TIS-CPC catalog requirements). Display capabilities include selection of records for display, sophisticated re-ordering facilities, and great flexibility in the selection and presentation of data for printing.

Unlike many data processing applications, where various kinds of processing take place on a regular basis once records have been added to a master file (for example, utility billing and accounting, where records are updated by consumption amounts, bills calculated and produced, payments recorded and balances computed, etc.), the Bibliographic/MARC Processing System carries out only information storage, retrieval, re-arrangement and summarization.

Although any individual record may be changed at any time, through the update process, there are at present no facilities for performing calculations or computations on data in records, other than simple tabulation of records within categories. The absence of facilities for computation, accumulation, statistical comparison, and similar intra-record or inter-record processing constitutes the systems's major specialization of function and hence unsuitability for this class of data processing applications.

Update programs

Construction, correction, and updating of the content of master file records is accomplished through the Update programs. These programs are:

1. Data Edit Program (BPSDATA) -- checks the document identification numbers of new records and corrections, and assigns an internal-format identification number for sorting purposes. Other features include automatic generation of multiple copies of transaction lines where these have been designated as belonging to more than one document identification number.

2. Sort Program (IBM Sort/Merge) -- sorts transactions into the same sequence as the master file. Multiple transactions for the same master file record are retained in the order in which they were received by the Data Edit program.

3. Update Program (BPSUPDT) checks transactions for a variety of error conditions, adds, changes, deletes, prints, and replaces records; adds, changes, deletes, and replaces fields within records; produces a proof sheet listing errors detected and the current contents of records or fields changed.

Extended features include a record selection capability (as an efficient procedure for selecting records whose did numbers are known), an extraction capability (to produce an abbreviated version of the file for special uses), and an edit capability (to change all occurrences of a specific data item throughout a range of records or the whole master file -- for example, changing the name of an organization following a merger).

Report generation programs

Selection, re-ordering, and display of information from the master file is accomplished through the Report Generation programs. These programs are:

1. Option Statement Preprocessor (QUERY) -- reads option statements describing specifications and requirements for selection, re-ordering and display, invokes standard option sequences from an option library, and passes option statements to the programs below.

2. Free Text Retrieval Program (BPSRET) -- selects records from the master file according to one or more sets of search criteria. Several independent searches can be carried out simultaneously. Record selection can be based on the presence or absence of particular variable fields, or on the content of variable or fixed field data. Compound selection rules can be formed from individual selection criteria and logical operators.

3. Indexed Thesaurus-based Retrieval Program (THESRET) -- selects records from the master file by making use of indexes to the file (generated in advance). Indexes may include a thesaurus, permitting automatic expansion of search terms to narrower (subordinate) categories. Because of the use of an indexed search procedure, file searching is much more efficient and economical with this program. However, the records must have fields suited for index generation.

4. Sort Key Edit Program (BPSSKED) -- prepares master file records (after selection as above or the entire file) for re-ordering by constructing a sort key according to parameters entered at execution time. The sort key is appended to the front of the MARC record to enable the file to be processed by the standard IBM Sort/Merge program.

The Sort Key Edit Program also generates additional copies of master file records where the same information is to appear at several places in a printed report (for example, a subject index would list the record under each of its subject terms). The facilities of this

Introduction

program and the following one enable master file information to be presented in ways very different from its original record organization.

5. Print Program (BPSPRT) -- prints data from master file records, with or without sort keys, according to specifications entered at execution time. Great flexibility is afforded for the content and positioning of report titles, page headings, and record data. Summary reports, tabulations, and indexes to previous reports may also be produced with this and the previous program.

Other programs

The Update and Report Generation programs constitute the basic methods of entering data into and displaying data from the system. Several other programs carry out special purpose tasks in these areas:

1. Key Word In Context (KWIC) Program (BPSKWIC) -- generates a KWIC listing of the data in fields selected at execution time. A KWIC listing contains one line for each word in each record (for the specified fields), in order by that key word: a powerful reference tool.

2. Thesaurus Manipulation programs -- produce hierarchical and alphabetic editions of a thesaurus encoded in MARC-format records. Minor programs provide thesaurus construction aids including facilities for file re-arrangement and renumbering.

3. Interrecord Data Transfer Program (BPSGEN) -- carries out systematic incorporation of data from "parent" records (e.g., journal title records) into subordinate records (issue and article records) to eliminate the need for repetitive entry of common information. The particular operations performed are those that have been defined by the needs of the TIS-CPC catalog maintenance operation.

4. Microfiche Formatting Program -- formats print lines generated by the Print program for production of microfiche by a private contractor. Fiche headings and master indices are generated.

5. File Inversion programs -- create direct access indexes to the master file to enable efficient, index-based searches via programs THESRET and TOBIAS (see below). Indexes can be generated for any variable field or fields in the record. If a thesaurus is available, an index is generated for it, permitting automatic expansion of search terms to include narrower (subordinate) terms if desired.

6. Interactive Subject Retrieval (TOBIAS) -- enables searches to be conducted on-line, in conversational mode. This program is being redesigned by the UNC-CH Institute for Research in the Social Sciences from a prototype developed by graduate students in the UNC-CH Department of Computer Sciences.