A review on several tools for conversions between different bibliographic formats

Shadi Khawandi, Anis Ismail, Firas Abdallah

Faculty of Technology, Lebanese University, Lebanon
ckh@ul.edu.lb
anismaiil@ul.edu.lb
firas.abdallah@gmail.com

***Abstract*—This paper describes the process of creating a schema for the conversion of MARC 21 (Machine Readable Catalog) bibliographic record, a cataloging format created by the Library of Congress, to UNIMARC (Universal Machine Readable Catalog) another format created by the International Federation of Library Associations and Institutions. This paper presents an overview of the history, techniques and importance of cataloging. It explains the problem being solved.**

***Keywords*— I MARC, bibliographic, Library, UNIMARC, cataloguing.**

I. INTRODUCTION

Nowadays cataloging is defined as the process of classifying or listing something into a catalog. In library and information science, the process includes the production of bibliographic descriptions of books as well as other types of discovery tools for documents. Today cataloging study and practice has broadened and merged with that of metadata ("data about data contents"), increasingly associated with Resource Description and Access [1].

The need for automating and standardizing the exchange of library cataloguing records have led the Library of Congress to commission several studies in the field of machine readable cataloguing. A pilot project called MARC, abbreviation for Machine Readable Catalogue or Cataloguing, resulted from these studies. The MARC format had the purpose to facilitate the creation and dissemination of cataloging records between libraries. The project was launched in the 1960s and later became by 1971 a US national standard for bibliographic data. This led to the creation of many formats of MARC around the world depending on the bibliographic recording standards in each country.

This revolution in the field of library cataloging was mostly due to the introduction of computer systems and new technologies into the library setting. The 20th century has many technological discoveries that affected directly the field of library research. Libraries all around the world have witnessed a large flow and exchange of information. Information and knowledge have become more and more accessible with the presence of contemporary communications and storage systems in addition to the availability of the internet.

By the late 19th century, an "information flood" occurred due to the availability of inexpensive printing. Library knowledge was organized by subject categories using card catalog systems. Currently cataloging the entire content of a library has become a task easily feasible using shared cataloging information. The tasks of exchanging, exporting or importing bibliographic records between libraries have become a daily common task. Nonetheless this task is somewhat difficult to achieve without the adequate tools of conversion between the different bibliographic records formats used by libraries.

In this paper, we present several tools publically for conversions between different formats. Among the types of conversions exists one that's not still widely researched, the conversion from MARC21 to UNIMARC (the two major MARC records format). Although the conversion from UNIMARC to MARC21 is already available, the opposite conversion is still not very well done or exists for very few library systems. There's no current open source tool that can perform the conversion appropriately. The current process of converting a MARC21 record into a UNIMARC record may involve using several tools; first to convert MARC21 to an intermediate format such as MARCXML or other formats and second to the UNIMARC format. The process may cause some loss of information during the conversion. The conversion itself from one format to another may require human intervention and correction to ensure the correct output. Which is a very intense task for catalogers considering the large number of records in each library.

Most of the Arab world and African nations do not use UNIMARC but use MARC21 instead or other standards for bibliographic records. With the increasing interest of the world in Arabic (and other foreign) publications and scriptures a new need surges; the conversion from MARC21 formats to UNIMARC. In addition UNIMARC is considered mainly a French standard it I's widely used in French libraries. Many Arab countries were historically colonized by France and are still under the influence of this colonization.

French colonies in Africa used French as an official language which also was used as the official educational system language. This fact greatly affected the ties between these countries and France in terms of politics and educational exchange. North African countries such as Tunisia and Algeria and some Middle Eastern countries such as Lebanon are among the consumers of French publications in the world.

II. Existing Solution

This paragraph presents the formats being used during the conversion. The formats are MARC21 and UNIMARC for bibliographic records. It explains briefly these formats structures; the tags, subfields and indicators used. It defines how they are created and their organization.

A. MARC21

The first source to visit concerning MARC specifications and conversion details is the Library of Congress. The Library of congress is as mentioned previously the pioneer in this study field. The library of Congress is still maintaining and updating the many uses of MARC formats and records depending on new needs and requirements of libraries.

The keystone for the development of automation in libraries was the simple but innovative MARC cataloging data format [2].MARC format was designed to overcome the difficult design requirements of cataloging or a catalog record itself. The most critical challenges data element and file length variability, large files, and update requirements were addressed with a simple format structure that embedded a directory to the data content fields in the front of the fields [2].

B. Content Designators

This directory is a simple partition of the catalog record into content designators that can be easily filled and accessed by any librarian familiar with the MARC format. The Library of Congress MARC21 summary statement of content designators as of april 2013[4] is presented in this paper.

C. Indicators

Indicators, two character positions follow each tag (with the exception of Fields 001 through 009). One or both of these character positions may be used for indicators. In some fields, only the first or second position is used; in some fields, both are used; and in some fields, like the 020 and 300 fields, neither is used. When an indicator position is not used, that indicator is referred to as “undefined” and the position is left blank. It is the convention to represent a blank, or undefined, indicator position by the character “#”.

Each indicator value is a number from zero to nine. (Although the rules say it can be a letter, letters are uncommon.) Even though two indicators together may look like a two digit number, they really are two single-digit numbers. The allowable indicator values and their meanings are spelled out in the MARC 21 documentation. In the example which follows, the first three digits are the tag (245 defines this as a title field) and the next two digits (a one and a four) are indicator values. The one is the first indicator; four is the second indicator.

Table 1
Example MARC Tag

245 14 $a The emperor’s new clothes / $c adapted from Hans Christian Andersen   and illustrated by Janet Stevens.

A first indicator value of 1 in the title field indicates that there should be a separate title entry in the catalog. In the card catalog environment, this means that a title card should be printed for this item and an entry for “Title” added to the tracings. A first indicator value of zero would mean that a title main entry is involved; the card would be printed with the traditional hanging indention, and no additional tracing for the title would be required (since it is the main entry).

Non filing characters: One of the more interesting indicators is the second indicator for the title field. It displays the number of characters at the beginning of the field (including spaces) to be disregarded by the computer in the sorting and filing process. For the title The emperor’s new clothes, the second indicator is set to “4” so that the first four characters (the “T,” the “h,” the “e,” and the space) will be skipped and the title will be filed under “emperor’s.”

D. Subfields

A subfield, most fields contain several related pieces of data. Each type of data within the field is called a subfield, and each subfield is preceded by a subfield code. Fields 001 through 009 have no subfields.

For example, the field for a book's physical description (defined by the tag 300) includes a subfield for the extent (number of pages), a subfield for other physical details (illustration information), and a subfield for dimensions (centimeters):

Table 2
Example Subfields for Book Physical Description

300 ## $a 675 p. : $b ill. ; $c 24 cm.

A subfield code: Subfield codes are one lowercase letter (occasionally a number) preceded by a delimiter. A delimiter is a character used to separate subfields. Each subfield code indicates what type of data follows it. (For each field in the MARC 21 bibliographic format, the MARC 21 documentation lists and describes the valid subfield codes.)

A delimiter: Different software programs use different characters to represent the delimiter on the screen or on printouts. Examples are a double dagger (‡), an "at sign" (@), a dollar sign ($), an underline (_), or the graphic symbol "‡". In this publication the dollar sign ($) is used as the delimiter portion of the subfield code.

In the example above, the subfield codes are $a for the extent, $b for other physical details, and $c for dimensions.

E. UNIMARC

Issued in 1977 by IFLA (International Federation of Library Associations and Institutions) UNIMARC (UNIversal MAchine-Readable Cataloguing) was primarily aimed at facilitating the exchange of bibliographic records originally produced in any other MARC formats. The design of UNIMARC was created to act as a common format capable of accommodating or translating data from/to other formats [5].

The main focus of UNIMARC was to facilitate the international exchange of records. This fact allowed UNIMARC to be recognized as a common standard by the Commission of the European Community (CEC) for data exchange among European national libraries, bibliographic utilities and the book trade. Following a workshop held in Luxembourg in 1990, a study on this matter was commissioned to the Deutsche Bibliothek whose results [6], highlighted the significance of UNIMARC for that purpose. The study also confirmed UNIMARC as the common standard for all European collaboration projects and focused on the need for data conversion programs from and to UNIMARC.

Over 36 years of existence, the UNIMARC format adapted to growing needs related to the coverage of new types of resources descriptions and changes in ISBDs (International Standard Bibliographic Description) such as ISBD(ER) [7], with concepts and terminology arisen from the new International Cataloguing Principles (ICP) and many other requirement related to the implementation of some functional requirements for cataloging standards. Generally, UNIMARC tag/indicator/subfield elements are represented as RDF properties, following ISBD practice. This document shows a complete list of fields, including obsolete and reserved fields [8].

III. CONCLUSION

In this paper, we described the two formats related to the subject of this study. It explains the process of creation of each format using the standard structure of the format and how cataloging records using these formats takes place.

MARC has served as an excellent medium for cataloguing and the exchange of cataloguing records for more than 4 decades. This was mainly due to the updates the formats had and is still having and to its adherence to international standards such as ISO 2709 through which bibliographic records can easily be transferred from one system to another. The UNIMARC format is also supported by the Z39.50 protocol. These two formats are the currently dominant formats around the world due to their high availability and ease of access.

REFERENCES

[1] K. a. D. H. Coyle, "Resource Description and Access (RDA): Cataloging rules for the 20th century.,"D-Lib Magazine, Jan./Feb. 2007.
[2] S. H. McCallum, "MARC: Keystone for Library Automation,"IEEE Annals of the History of Computing, pp. 34-49, April–June 2002.
[3] J. Knapp, "Design Considerations for the MARC Magnetic Tape Formats,"Library Resources & Technical Services, vol. 12, pp. 275-284, January 1968.
[4] L. o. Congress, "Library of Congress," April 2013. [Online]. Available: http://www.loc.gov/marc/bibliographic/bdsummary.html. [Accessed 01 April 2013].
[5] R. M. Galvão, "UNIMARC – Understanding the past to envision future," in New futures for bibliographic data formats: reflections and directions — UNIMARC Core Activity , Lisbon, Portugal, 2012.
[6] "Study to establish the feasibility of using UNIMARC amongst EC national libraries, bibliographic utilities," Luxembourg, 1992.
[7] A. M. Sandberg-Fox, "The ISBD(ER) and New Developments in Cataloging Electronic Resources," Fairfax, Vermont.
[8] A. Hopkinson, 2008. [Online]. Available: https://eprints.mdx.ac.uk/2655/1/UNIMARC-2008-AH_final.pdf.

Indeed, several competing formats heterogeneity hinders the design and development of interoperable and standards for the representation, storage and exchange of systems and entails critical integration issues for the Health Care Information Systems. ECG recordings can be found in the digital ECG standardization Objectives: This article aims at performing a comprehensive arena. Such formats and standards have successfully been overview on the current state of affairs of the interoperable applied in prototypes or even real environments, as has been exchange of digital ECG signals. This includes