Papers of BAS
Humanities and Social Sciences
Vol. 4, No 1, 2017

LIBRARY SCIENCE

MACHINE-READABLE FORMATS AND THE PROBLEM OF CONVERSION OF DATA BETWEEN SYSTEMS FOR BIBLIOGRAPHIC INFORMATION IN THE FIELD OF HUMANITIES (ON MATERIAL FROM THE CYRILLO-METHODIAN BIBLIOGRAPHY)

Nelly Gancheva

Abstract: The history of formats for bibliographic information from their emerging in the 60s of the 20th century until today is presented. The specific features of the two main machine-readable formats for data exchange – MARC (MARC 21) and UNIMARC are put forward, as well as their application in Bulgaria. On this basis the processes of automation of the Cyrillo-Methodian bibliography are assessed – the initial choice of software, the following updates, the reasons for data conversion from the medium of CDS/ISIS into a new information environment – that of MARC 21. Emphasis is placed on the problems following the conversion. Their presence confirms the thesis on the need for global unification and standardisation of the electronic presentation of metadata.

Keywords: CDS/ISIS, formats for bibliographic information, ISO 2709, MARC (MARC 21), UNIMARC, Cyrillo-Methodian bibliography, conversion.

The automation of library and bibliographic activities internationally is a process that began in the 60s of the 20th century. It is associated primarily with the creation of the machine-readable format for bibliographic description MARC (Machine-Readable Cataloguing). Machine-readable cataloguing means that data from the catalogue record can be read and interpreted by the computer. Raising the technological level where the bibliographic information exists led to the development of the so-called "information matrix" for bibliographic description of documents. This matrix is embedded in the international standard ISO 2709 approved by the International Organization for Standardization (ISO), according to which elements of the description are grouped into well-defined fields and subfields. Compliance with this standard guarantees unhindered exchange of bibliographic information, structured in different formats and processing based on different information systems. Created in the

70s of the 20th century first International Standard for Bibliographic Description (ISBD) represents a favourable environment for the development and construction of the machine-readable formats because standardising the fields and structure of the description allows it to be easily represented by fields of formats for machine-readable records. Over the next decades international standards for bibliographic description are constantly evolving, pass through many edits, periodically their updated versions are published, which are the fruit of efforts towards the constant renewal and adaptation to the rapidly changing information environment.

Machine-readable formats for bibliographic records

1. MARC

The USA Library of Congress is the organisation that assumes the leadership of the project for creation of a machine-readable format for bibliographic records. The project initially involved 16 US academic and scientific libraries. Team leader in the development of MARC is the specialist in programming and systems analysis Henriette Avram. After processing the pilot project the format is put into use in 1967.

The main purpose of the format is to allow free exchange of records between databases built on the basis of different systems. The structure of the MARC format is standardised by the American National Standards Institute (ANSI), Standard Z39.2. It defines the type of the identification codes of the elements of the record. The content of the record is determined by standards (rules) for bibliographic description, which are not related to the format itself, but to the data needed to identify the types of documents.

2. ISO 2709

International standard for data exchange, which determines the structure of records.

3. UNIMARC

Universal communicative format that can be used not only for converting data from one format to another, but also as a basic format for creating bibliographic records and building databases. UNIMARC is being supported by the Program for Universal Bibliographic Control and the International MARC, developed by IFLA (International Federation of Library Associations and Institutions). The first edition of the format was in 1977, followed by numerous amendments that updated and completed the content.

A significant number of the national bibliographic agencies adopted MARC as a national format for bibliographic records, others preferred the specifics of UNIMARC. This greatly hinders international data exchange. Indeed, these formats have much in common in so far as they are built on MARC. Between them, however, there are many significant differences arising from the specifics of cataloguing in individual countries.

In the 90s of the last century, the European Union recognised UNIMARC as an official format for the exchange of data between Member States of the Union. So it became the only internationally recognised communicative format developed under the auspices of IFLA.

4. MARC 21

At the beginning of the 21st century MARC 21 appeared – designed to be used as a common format for bibliographic data of the USA, Canada, UK, Australia and New Zealand. Soon after its creation, it began to spread rapidly even outside these countries, as most of the producers of library software prefer it when creating electronic catalogues and databases.

The existence of two main formats for data exchange – UNIMARC and MARC 21 – causes profound discussions which one is more suitable for the needs of libraries, information centres and bibliographic agencies. Such discussions exist in our country.[1] Undoubtedly more internationally widespread is MARC 21. This however does not predetermine the choice of format because there are many other factors that influence it. The comparative analysis between UNIMARC and MARC 21 outlines the following in their specifics. Both formats structure data types into separate groups called blocks. The main difference between them is that they are bound by different international standards of bibliographic description. While UNIMARC is fully consistent with the structure of ISBD, MARC 21 is not so closely tied to the ISBD. MARC 21 is consistent with the individual national cataloguing rules, and in particular with the second edition of the Anglo-American Cataloguing Rules (AACR 2). Moreover UNIMARC provides opportunities to create richly structured records with flexible connections between them, and in MARC 21 the structure of records is permanently established and offers no additional opportunities to change it. UNIMARC is being characterised by the presence of multiple binding mechanisms and provides choice of methodology for creating records on many levels. These differences are the reason for the lack of compatibility between the formats in some of the records and the difficulties in transferring them from MARC 21 to UNIMARC, and vice versa. Generally the conversion from one format to another is difficult, especially for records on many levels, which requires them to be edited manually one by one. It is also important to note that even when bibliographic systems use the same formats, if terms of description are different, this would lead to differences in the records of identical documents and would cause considerable difficulties in exchanging information.

The above characteristics of UNIMARC make it attractive for a significant part of the national bibliographic agencies in Europe, which enter it as a basic format on a national level. National Library of Bulgaria also adopted UNIMARC – cataloguing documents there is made in the format COMARC, which is a subset of UNIMARC.

---

[1] Detailed analysis of the current state of cataloguing in Bulgaria in the context of the processes worldwide, was made by M. Milanova in her dissertation for awarding the educational and scientific degree "doctor" [see **Milanova, 2008**].

In recent years we have witnessed the successful use of MARC through the markup language XML (Extensible Markup Language). Now there are valid versions of XML that are in agreement with MARC. They are presented as international standards, which are exempt from many conventions of the record structure in MARC, yet retain some of the important characteristics of the format – labels, subfields and indicators. As an example we can point out the developed by the USA Library of Congress product MARCXML, where a successful conversion of MARC into an XML structure is implemented.

Automation of Cyrillo-Methodian Bibliography

The brief overview of internationally accepted formats for bibliographic record for us has the function of an information basis on which to build upon, evaluating the process of automation of Cyrillo-Methodian bibliography – the main subject of this presentation.

The first bibliographic reference book in the field of Cyrillo-Methodian studies is work of Grigoriy Ilyinsky and was released in 1934 as an edition of the Bulgarian Academy of Sciences [**Ilyinsky, 1934**]. Over the next decades three more general Cyrillo-Methodian bibliographies were published [**Popruzhenko, Romanski, 1942; Mozhaeva, 1980; Duychev, Kirmagova, Paunova, 1983**]. Since 1980 the collection, processing and issuance of Cyrillo-Methodian bibliography has been carried out by the Cyrillo-Methodian Research Centre at the Bulgarian Academy of Sciences (CMRC). It continued the activity of the restored in 1971 Cyrillo-Methodian Commission of BAS, which marked the beginning of the systematic collection of this bibliography. The Cyrillo-Methodian Research Centre is the only institution that prepares bibliography of the global scholarly and popular literature of the Cyrillo-Methodian studies, as well as of artistic works (literary, musical, works of art, etc.) dedicated to the Cyrillo-Methodian work. The Centre maintains an electronic database containing bibliographic records of about 30 000 publications issued after 1940. There are also two bibliographies prepared in the Centre, which contain descriptions of Bulgarian publications for the period 1846-1944, which are not reflected in the bibliographies of Gr. Ilyinsky and M. Popruzhenko-St. Romanski [**Zhelyazkova, Zafirova, 2003; Zhelyazkova, Zafirova, 2010**].

In 2000 the bibliographic database of CMRC was built. The software product **CDS/ISIS** (Computerized Documentation System – Integrated Set for Information Systems), maintained and distributed by UNESCO, was selected for the automation. At that time there were really no big choices – CDS/ISIS has been the only product approved on a national level, consistent with the structure of the machine readable formats. Therefore, the automation of our national bibliography was carried out on its basis. This product has the following features: input and output format ISO 2709, which regulates the structure of the machine-readable bibliographic record; versions are adapted in Cyrillic; it is distributed and improved for free. In the last decade of the 20th and the beginning of the 21st century these several characteristics satisfied the requirements of libraries and bibliographic centres in Bulgaria and the overall assessment of CDS/ISIS was positive.

Since the beginning of the 21st century until nowadays quite naturally several changes in the informatisation process of cataloguing worldwide occurred.

Permanent development of formats and standards for bibliographic description and bibliographical information systems exerted its influence on the processes in the country, as well. In Bulgarian libraries two formats apply for bibliographic record – UNIMARC (National Library) and MARC 21 (in libraries, united in the National Academic Library and Information System, NALIS). For incorporation of formats different software products are used – such that are developed or adapted in our country as well as foreign software packages for library automation (ALEPH, Q-Series, VTLS). Using different software products, on the one hand and the need to realise data exchange in national and international framework, on the other hand, requires adequate solutions to ensure access to information and convenience for the user. Unfortunately, we are witnessing some disturbing phenomena that can have broad and lasting negative effect.[2] Due to insufficient knowledge of the philosophy of formats in Bulgarian libraries there is violation of the achieved standardisation of bibliographical description. Individual libraries that have implemented and use MARC 21 format, introduce Anglo-American rules for bibliographic description in records they create. They do not comply with national standards and do not understand the difference between the format for data exchange that defines the structure of the record and a standard for bibliographic description that defines what different structural format fields will contain. This alarming finding shows that it is necessary formats to be further studied, analysed and presented to the library community, which is a prerequisite for their understanding and correct implementation and is the basis for making an informed decision when choosing them.

Cyrillo-Methodian bibliography in a new information environment

Changes in the field of cataloguing in the country provoked a change in the bibliographic database of CMRC, as well. An analysis of the capabilities of the software product used to create the database showed that it is currently obsolete and modifying it would not be effective, which called for it to be replaced by another. When choosing a machine-readable format for records we chose MARC 21. This decision was made without hesitation for the following reasons. First, the structure of bibliographic records in MARC complies with the requirements of standard ISO 2709. Second, the academic libraries in Bulgaria work with MARC 21. Third, the MARC 21 format is a successful one, with a good outlook. By adopting it for the first time it became necessary the structure of the Cyrillo-Methodian bibliography records to comply with the requirements of a particular format. When choosing a partner for implementing the project of transferring the database to a new information environment we focused on cooperation with the Central Library of the Bulgarian Academy of Science (CL BAS) and NALIS, which work with MARC 21 in the environment of the integrated library system ALEPH 500.

In fulfilment of this decision, transfer of all bibliographic records of the electronic database maintained by CMRC to the platform ALEPH 500 of

[2] These phenomena are thoroughly examined in the dissertation of M. Milanova [see **Milanova, 2008:** 134-140].

CL BAS and their integration into the Union Catalogue of NALIS was prepared and carried out. The Cyrillo-Methodian bibliography is presented there as an independent resource, copyright holder of which is CMRC.[3]

The original forecast that with the conversion of data many difficulties will arise and many problems will have to be solved, confirmed fully in the work process. This forecast was based on the fact that despite the existence of a relationship between CDS/ISIS and MARC 21 arising from compliance with common standards, there are many significant differences between them. In comparison with the structure of CDS/ISIS the structure of the MARC 21 format is more complex, data is grouped in strict hierarchical order, a number of conventions have to be observed upon their entry. In MARC data types are structured into separate groups called blocks. Information for each entry is placed in the following departments: 1. record marker or leader containing data about the structure of the record, its status and type, about its completeness, and the rules by which bibliographical description is made; 2. directory; 3. fields marked with three-digit numbers, called tags, which are followed by indicators and subfields in which data on the relevant elements of the description are recorded (see Application 1).

Application 1. MARC 21. Fields for basic information – author and title

Field 100. Personal name / author main entry
Indicator 1. Type of personal name
0 – First name;
1 – Surname;
3 – Coauthors with the same surname.

Indicator 2 – Fell out of use in 1990.
Subfields:
$a – Personal name
$q – Full form of the name
$d – Dates

Example

100 1# $a Kuev, Kuyo M.
$q (Kuyo Kuev)
$d 1909–1991

Field 245. Title

Indicator 1. Title added entry:

0 – Author is not specified;
1 – Author main entry, title added entry.

Indicator 2: Ignoring characters (0 to 9) – the number of characters that are not taken into account in alphabetical order; usually labelled with 0.

Subfields:

$a – Main title
$b – Subtitle, parallel title etc.

[3] Kirilo-Metodievska bibliografiya. Available from: aleph.cl.bas.bg/F/EJMDBSGCAE-91GG28D2VUMMB9817HT32AGVI9GBGQJQCT96GALD-00065?func=find-b-0&local_base=KMNC [accessed 4 November 2016].

Example

245 10 $a Sadbata na starobalgarskata rakopisna kniga prez vekovete

Based on comparative tables between the fields in CDS/ISIS and MARC 21, an analysis of the structures of both types of records was made. The analysis helped to develop conversion programs that were used to transfer information from the environment of CDS/ISIS into the environment of MARC 21. The conversion was carried out in two stages. The first involved transfer of CDS/ISIS records into ISO format, the second – converting the ISO file into MARC through conversion programs (see Application 2).

Application 2. A comparative table between the main fields in the structure of CDS/ISIS "Cyrillo-Methodian bibliography" database and MARC 21

Catalogue Fields	CDS/ISIS field number*	MARC 21 field number
ISBN	080	020
Author	015	100
Title	050	245
Edition statement	052	250
Editor	151	700
Publishing data	042/056	260
Physical description	051/053/058	300
General note	070	500
Subject	160	650

* It is presented numbering of the fields of CDS/ISIS database "Cyrillo-Methodian bibliography".

As it is evident for some of the areas of bibliographic description in the CDS/ISIS database several fields are provided. For example, publishing data is entered in two fields (042 and 056) and the physical characterisation of documents – in three (051, 053 and 058). For the same data, however, MARC 21 provides only one field – field 260 respectively for publishing data and field 300 for physical characteristics. This example demonstrates just one of the many problems of conversion associated with targeting information from several fields into one field. Despite the measures taken in the new information environment many entries (mainly of articles) contain identical errors due to the discrepancies between the fields of both products, which are not subject to automatic removal. This required them to be manually corrected.

Conversion of data in the Cyrillo-Methodian bibliography is another example of the difficulties accompanying the transfer of bibliographic records from one information environment into another. Experience in this case confirms the correctness of the thesis on the need for global unification and standardisation of the electronic presentation of metadata.

List of abbreviations

AACR – Anglo-American Cataloguing Rules
ANSI – American National Standards Institute
CDS/ISIS – Computerized Documentation System – Integrated Set for Information Systems
CL BAS – Central Library of the Bulgarian Academy of Science
CMRC – Cyrillo-Methodian Research Centre at the Bulgarian Academy of Sciences
IFLA – International Federation of Library Associations and Institutions
ISBD – International Standard for Bibliographic Description
ISO – International Organization for Standardization
MARC – Machine-Readable Cataloguing
NALIS – National Academic Library and Information System
UNIMARC – Universal Machine-Readable Cataloguing
XML – Extensible Markup Language

REFERENCES

Duychev, Ivan, Angelina Kirmagova, Anna Paunova. 1983. *Кирилометодиевска библиография. 1940–1980* [Kirilometodievska bibliografiya. 1940–1980]. Sofia: SU "St. Kliment Ohridski".

Ilyinsky, Grigoriy. 1934. *Опыт систематической кирилло-мефодьевской библиографии* [Opyt sistematicheskoy kirillo-mefodyevskoy bibliografii]. Sofia: Balgarska akademiya na naukite.

Milanova, Milena. 2008. *Българската каталогизация в глобалното информационно пространство на XXI век. Анализи, стратегии, перспективи.* Дисертация за присъждане на образователната и научна степен „доктор" [Balgarskata katalogizatsiya v globalnoto informatsionno prostranstvo na XXI vek. Analizi, strategii, perspektivi]. Sofia: SU "St. Kliment Ohridski". Available from: research.uni-sofia.bg/handle/10506/1089 [accessed 4 November 2016].

Mozhaeva, Inessa. 1980. *Библиография по кирилло-мефодиевской проблематике. 1945–1974 гг.* [Bibliographiya po kirillo-mefodievskoy problematike. 1945–1974 gg.]. Moskva: Nauka.

Popruzhenko, Mihail, Stoyan Romanski. 1942. *Кирилометодиевска библиография за 1934–1940 год.* [Kirilometodievska bibliografiya za 1934–1940 god.]. Sofia: Balgarska akademiya na naukite i izkustvata. Kirilometodievska komisiya.

Zhelyazkova, Veselka, Nedyalka Zafirova. 2003. Българска кирило-методиевска библиография. 1846–1934 г. [Balgarska kirilo-metodievska bibliografiya. 1846–1934 g.]. In: Nikolova, Svetlina. (Ed.). *Kirilo-Metodievska bibliografiya. 1516–1934*. Sofia: Kirilo-Metodievski tsentar, 419-685.

Zhelyazkova, Veselka, Nedyalka Zafirova. 2010. Българска кирило-методиевска библиография. 1935–1944 г. [Balgarska kirilo-metodievska bibliografiya. 1935–1944 g.]. In: Nikolova, Svetlina. (Ed.). *Kirilo-Metodievska bibliografiya. 1934–1944*. Sofia: Kirilo-Metodievski tsentar, 203-389.

Correspondence address:
Nelly Gancheva, Chief assistant professor, PhD
Cyrillo-Methodian Research Centre
Bulgarian Academy of Sciences
13 Moskovska Str.
1000 Sofia, Bulgaria
Tel. (359) 29870261
E-mail: nelly_vasileva@abv.bg