Cataloguing dataset in Library Information Systems using the MARC 21 format

Rudić Gordana*, Ivanović Dragan**
* University of Novi Sad, Faculty of Sciences, Novi Sad, Serbia
** University of Novi Sad, Faculty of Technical Sciences, Novi Sad, Serbia

goga@dmi.uns.ac.rs, dragan.ivanovic@uns.ac.rs

Abstract— Besides cataloguing various kinds of publications, a library information system (LIS) can also catalogue supplement materials including dataset. This is especially important if the LIS is used at an academic institution, where the institution librarians should also catalogue scientific outputs of its researchers. This paper presents how dataset can be catalogued in library information systems based on worldwide used bibliographic format MARC 21.

I. INTRODUCTION

Library information system (LIS) catalogues various kinds of publications including supplement materials and dataset. Cataloguing of research dataset is especially important if the LIS is used at an academic institution, where the institution librarians also catalogue scientific outputs of its researchers. Cataloguing dataset within LIS will enhance the reuse of scholarly data and development of Open Science. A diverse set of stakeholders representing academia, industry, funding agencies, and scholarly publishers designed a set of principles referred as the FAIR Data Principles, i.e., a guideline for those wishing to enhance the reusability of their data holdings. Implementation of dataset cataloguing within LIS is in accordance with those principles.

The MARC 21 formats are standards for the representation and communication of bibliographic and related information in machine-readable form. The formats are maintained by the Library of Congress in consultation with various user communities. There is a lot of library information systems (LIS) worldwide which are based on MARC formats including BISIS – a LIS developed and used at University of Novi Sad. Three versions of BISIS LIS were based on UNIMARC format. Version 4 includes its own editor for creating bibliographic records in UNIMARC and MARC 21 [17, 18], and mapping from UNIMARC to MARC 21 was defined [19]. The motivation for dataset cataloguing in the BISIS LIS is improvement of reusability of researchers’ data.

Various aspects of cataloguing dataset are main topic of a set of scientific publications. In the paper [21] authors represent creating a “data catalog” that describes different datasets, big population-level studies, as well as small datasets created by researchers. Authors of the paper [23] propose an approach to describe security-related datasets by means of a Dataset Description Language (DDL) and a Dataset Description Sheet (DDS). Online catalog of research data catalog, Databib [22], is a tool and resource for locating repositories of research data. Librarians and other information specialists identify and catalogue different data repositories using Double Core, Friend of a Friend and Databib Terms. Although some approaches for cataloguing research dataset already exist, those approaches cannot be applicable for the specific need of University of Novi Sad. Academic LISs are very often interoperable with institutional repositories, virtual research environments and research information systems. Besides BISIS library information systems used by University of Novi Sad libraries, research management system at University of Novi Sad (CRIS UNS) has been also developed on a MARC 21 compatible data model in order to enable an easy integration of research management system and academic LIS at faculties of University of Novi Sad. The research question is whether it is possible to catalogue research dataset in the MARC 21 format and use those MARC 21 records to describe research dataset for the sake of library information systems and the CRIS UNS research management system. Authors of paper [20] explain the process and discuss advantages of cataloging the data and adding records for the data to their integrated library system based on MARC 21.

CRIS UNS is the main result of the DOSIRD UNS project (http://dosird.uns.ac.rs/). The CRIS UNS system is built on CERIF compatible data model based on the MARC 21 format. MARC 21 is used for representing authority data as defined in CERIF (authors, institutions, projects, and events). Furthermore, MARC 21 is also used for representing data about published results. This approach provided consistent handling of authority and bibliographic data and the support for multilingual content. The thorough use of MARC 21 facilitates data exchange with CERIF and MARC-based systems including BISIS LIS, comprehensive bibliographic reports, and evaluation of research results. We would also like to use MARC 21 format for cataloguing datasets.

II. DOSIRD UNS

The first faculties in Novi Sad were founded in 1954. The University of Novi Sad was founded on 28th of June 1960 and it represents an autonomous institution for education, science and arts. While software infrastructure for educational domain of the University of Novi Sad is well developed, there is lack of software infrastructure for science and arts. The main goal of the DOSIRD UNS project is developing of software infrastructure for research domain of the University of Novi Sad. The project has been started in 2009.

CRIS UNS (Current Research Information System at University of Novi Sad) system is under development since the year 2009 [1, 2] and it is the main result of the
DOSIRD UNS project. It has been implemented with an intention to fulfill all specific requirements prescribed by rule books of the University of Novi Sad, Provincial Secretariat for Science and Technological Development of Autonomous Province of Vojvodina, and Ministry of Education, Science and Technological Development of Republic of Serbia. In addition, the substantial requirement to be fulfilled by the system CRIS UNS is its interoperability with systems possessing large databases of scientific research results, thus improving the visibility of the scientific results achieved by researchers from University of Novi Sad and therefore raising the rating of the University.

Informational requirements imposed to this system are:

1. Access to the application via an arbitrary modern web browser.
2. Researchers entering their references by themselves without the need to be at home on any standard for references describing.
3. Interoperability of our system with diverse systems containing scientific content such as CRIS systems, institutional repositories, library information systems, digital repositories, etc. [4, 10, 11, 12]
4. Capability to search the database containing scientific results [3, 9].
5. Capability to perform evaluation of scientific research results following the rule book(s) prescribed by the Ministry of Education, Science and Technological Development of Republic of Serbia [13, 14, 15].
6. Reporting for the faculties, University, Provincial Secretariat for Science and Technological Development of Autonomous Province of Vojvodina and Ministry of Education, Science and Technological Development of Republic of Serbia [16].

PHD UNS is digital library of dissertations defended at the University of Novi Sad since its foundation. It has been developing since 2010 at University of Novi Sad, Serbia. The first version was put into operation in 2013. The digital library of dissertations is a web application implemented using Java platform and set of open-source libraries written in Java. Data about dissertations are stored in MARC 21 format [7, 8]. Architecture of the digital library enables an easy integration with library information system as well as easy integration with research information system based on CERIF standard [5].

The PHD UNS library stores the entire process of a dissertation’s submission and defense. The digital library is integrated with the CRIS UNS system in order to create a unified central catalog of all scientific-research outputs published by the University

III. MARC 21

The MARC 21 formats (www.loc.gov/marc) are standards for the representation and communication of bibliographic and related information in machine-readable form. The standard consists of five documents: MARC 21 Format for Bibliographic data, MARC 21 Format for Authority Data, MARC 21 Format for Holdings Data, MARC 21 Format for Classification, MARC 21 Format for Community Information. The formats are maintained by the Library of Congress in consultation with various user communities. As it is already stated, there is a lot of library information system (LIS) worldwide which are based on MARC formats.

A MARC 21 record starts with the leader which is the first field; it has fixed-length, and contains 24 character positions where the character positions specified the meaning of data (code) stored on that position. After the leader, the MARC 21 record lists control fields which also have fixed-length and character positions. After the list of control fields, the MARC 21 record has a list of data fields which have different structure. Data fields consist of a maximum of two indicators and a limited set of subfields, which are the carriers of information.

The MARC 21 format for bibliographic data is designed for cataloging different types of material like textual material, computer files, cartographic, music material, as well as datasets. Establishing the link between bibliographic records describing different kinds of material is possible using the linking entry fields [6]. We suggest linking of certain scientific publication with related dataset in that way.

IV. DATASET METADATA FORMATS

A research results presented in some publication can be based on a dataset. Dataset should be catalogued and available for society for at least two reasons. The first one is openness of science which accelerates further development of knowledge based society. Some new research can be conducted using the same dataset which has been created in some previously published research. The second reason is transparency of science. Each published research should be repeatable and verifiable. If some research publication describes methodology, dataset, and results of research, and dataset is not available for downloading, then someone cannot repeat the same research to check the results. This is especially important taking into account that a lot of published researches are funded by local government or European Union funding programs. Besides preservation of dataset digital files, it should be also catalogued using some metadata format.

DCAT is an RDF vocabulary recommended by W3C (https://www.w3.org/TR/vocab-dcat/). It is designed to facilitate interoperability between data catalogs published on the Web, and defines three main classes (Figure 1):

- dcat:Catalog represents the catalog
- dcat:Dataset represents a dataset in a catalog.
- dcat:Distribution represents an accessible form of a dataset

In addition, specific application profiles for DCAT are being developed that are relevant for the use of DCAT in certain domains:

- DCAT-AP is a profile based on DCAT for describing public sector datasets in Europe.
- GeoDCAT-AP is an extension of DCAT-AP for describing geospatial datasets, dataset series, and services.

CKAN (Comprehensive Knowledge Archive Network) is a web-based system for the storage and distribution of open data. Furthermore, CKAN makes data discoverable and presentable, and provides a web page with a rich collection of metadata for each dataset, making it a
valuable and easily searchable resource. The central entity in CKAN data model is the Dataset entity which can be linked with the Resource entity (actual data: files, APIs etc.) and with set of metadata. CKAN supports a variety of “core” metadata for a dataset, but it can be even extended with an unlimited amount of arbitrary additional metadata in the form of “extra” key/value associations. Datasets can be grouped and linked to each other.

![Figure 1. DCAT](image)

V. CATALOGUING DATASET METADATA IN THE MARC 21 FORMAT

If someone wants to catalogue a dataset in the MARC 21 bibliographic format, it is necessary to determine the type of record that will be described, bibliographic level, file type, etc. Type of a record in MARC 21 format is prescribed by the 6th position of a record leader. The record type that describes a dataset can be determined by the code m – which means that the certain MARC 21 record presents a computer file. The bibliographic record level determines the code at the 7th position of the record leader. If the dataset is an integrative information source, then the level of bibliographic record is determined by the code i. The file type is determined by the code at the 26th position of the control field 008 (for example, a - numerical data, e - bibliographic data, j - online system or service, etc.). The record identifier is stored into the control field 001. The physical characteristics of the dataset are stored into the control field 007. The category of material should be stored at zero position of the field 007, for example, the code c represents electronic source of information. The other positions of the field 007 should be used for cataloguing information about can be entered specific material designation (o - optical disk, r - distance source, s-standalone device, etc.), dimensions, image bit depth, compression level, etc. The dataset title should be stored in the subfield a of the data field 245 (245a), the dataset publisher should be stored in the subfield 260b, the dataset URI should be stored in the subfield 856u, etc.

For instance, if we have the following example of a dataset metadata in DCAT (Listing 1), someone can represent that in the MARC 21 format (Listing 2).

```
Listing 1. A dataset metadata in DCAT

dct:title "Dataset title" ;
dct:issued "June 14, 2013, 4:36 a.m." ;
dct:modified "June 14, 2013, 4:36 a.m." ;
dct:publisher "Name of publisher";
dct:format "Format of digital file";
dcat:keyword "Subject of dataset";
dcat:mediaType "Type of digital file";
dcat:accessURL
<https://some.domain.eu/someExcelDocument.xls> ;
dcat:size [dcat:bytes "22528"]
```

```
Listing 2. A dataset metadata in MARC 21

001  00002222
245[a] Dataset title
260[b] Name of publisher
347[a] Type of digital file
347[b] Format of digital file
347[c] 22528 bytes
518[a] Created June 14, 2013, 4:36 a.m.
518[a] Modified June 14, 2013, 4:36 a.m.
650[a] Subject of dataset
856[u]
https://some.domain.eu/someExcelDocument.xls
```

VI. LINKING MARC 21 RECORDS REPRESENTING DATASETS AND SCIENTIFIC PUBLICATIONS

Links between a MARC 21 record representing a dataset and other MARC 21 records representing scientific
outputs based on the certain dataset can be established using chronological, vertical or horizontal links.

If the certain dataset was used for a thesis research, and the dataset is supplement of the thesis, then the link can be established by the data field 772 in the MARC 21 record of the dataset. The data field 772 should contain information about the MARC 21 record of the thesis. For instance, if the dataset shown in Listing 2 was used in a thesis represented by the MARC 21 record with identification number 00001111, the thesis title is Thesis Title, and the thesis author is Thesis Author, then the Listing 2 should be extended with the following data field:

772 [i] Supplement to:
[a] Thesis Author
[t] Thesis Title
[w] 00001111

Conversely, the MARC 21 record of the thesis in the data field 770 should store information about the used dataset:

770 [i] Has supplement:
[t] Dataset title
[w] 00002222

The shown example presents a vertical relationship between the MARC 21 records.

Moreover, a MARC 21 bibliographic record of dataset can be linked with MARC 21 authority records. MARC 21 authority records should be a carrier for information concerning the authorized forms of names and subjects. In this way, names of persons or institutions responsible for creation and maintenance of dataset can be authorized.

VII. CONCLUSION

An approach for cataloguing dataset using MARC 21 format was presented in this paper. Academic library information systems should be interoperable with institutional repositories, virtual research environments and research information. Thus, library information system (BISIS) and research management system (CRIS UNS) developed at University of Novi Sad are based on MARC 21 format in order to enable an easy integration of research management system and academic LIS at faculties of University of Novi Sad. The paper presents how a DCAT representation of dataset can be transformed to MARC 21 format, and how dataset MARC 21 record can be linked with scientific publication MARC 21 record.

Further research includes extension of CRIS UNS system with a component for dataset management. Although MARC 21 format will be used for describing datasets, the dataset management component should enable researchers from University of Novi Sad to catalogue and preserve their datasets to the CRIS UNS system without knowledge of the MARC 21 format.

ACKNOWLEDGMENT

The work is supported by Ministry of Education, Science and Technological Development of the Republic of Serbia, through project no. III47003: “Infrastructure for Technology Enhanced Learning in Serbia”.

REFERENCES


