Mapping scheme from Invenio to CERIF format

Valentin Penca*, Siniša Nikolić *, Dragan Ivanović *

* University of Novi Sad/Faculty of Technical Sciences/Department of Computing and Automatics, Novi Sad, Serbia
{valentin_penca,sinisa_nikolic, chenejac}@uns.ac.rs

***Abstract*—This paper describes basics of the Invenio institutional repository and CRIS systems and their data models. The result of this research is mapping scheme of the data from Invenio to the CERIF standard.**

I. INTRODUCTION

Nowadays, there is a trend for knowledge dissemination which consequently includes preservation of publication and accompanying resources. Therefore, one of the most important tasks is how to preserve and make that data accessible and interoperable. One of the most important tasks is how to preserve and make that data accessible. Institutional Repository (IR) can solve the mentioned issue. In [1], an IR is addressed as an electronic system that captures, preserves and provides access to the digital work products of a community. The three main objectives for having an institutional repository are:

1. creating global visibility and open access for an institution's research output and scholarly materials.
2. collecting and archiving content in a "logically centralized" manner even for physically distributed repositories.
3. storing and preserving other institutional digital assets, including unpublished or otherwise easily lost ("grey") literature.

Firstly, there had been only several implementation of the IRs with limited set of features, which later became useful and world-wide popular tools such as Digital commons and ContentDM [2]. It is true that improvements in the field of WEB technologies contributed to rapid development of IRs around the world. The availability of open-source technologies affect on the rapid development of IRs worldwide, particularly among academic and research institutions. Thus, it is not surprising the existence of several open-source software platforms available for developing IRs such as *Invenio* [3]. *Greenstone* (GS) [4], *EPrints* [5], *DSpace* [6], *Fedora* [7] and *SobekCM* [8]. Although IRs had been used for a long time, they still don't have a mutually agreed and standardized representation of their data. Consequently, that can cause difficulties in data exchange between diverse IR systems.

So, to overcome these problems in the data exchange, one of the possible solutions is to rely on some predefined standard outside the IR field. Common European Research Information Format (CERIF) standard [9], which is the basis of Current Research Information Systems (CRISs), is used for data exchange from scientific-research domain and can be utilised in IR domain. It is reasonable to assume that some mappings of the IRs data to the CERIF model needs to be proposed in order to solve the problem of the data exchange among diverse IRs.

In this paper the scheme for mapping data from Invenio IR to CERIF format is proposed. Specifically, this paper will address the mapping of different kinds of publications and other recourses that are related to them from the Invenio to CERIF model. That scheme can be used as a guideline, supporting the exchange between Invenio repositories and CRIS systems. Motivation for this work was also to extend and improve research from [10] [11] [12].

II. INVENIO IR

The Invenio is one of the most popular IR solutions and it is used by the CERN and other research institutes which are outside the CERN such as SLAC National Accelerator Laboratory, Fermilab, and the École Polytechnique Fédérale de Lausanne. The Invenio is a free software suite enabling client to run his own digital library or document repository on the web. The technology offered by the software covers all aspects of digital library management, from document ingestion through classification, indexing, and curation up to document dissemination.

The Invenio is a completely an open source software library management package that provides the tools for management of digital assets in an institutional repository. The software is typically used for open access repositories for scholarly and/or published digital content and as a digital library. Actually, it is a free software licenced under the terms of the GNU General Public Licence (GPL). This provides a straight-forward advantage for institutions with smaller budgets, that have programmers on their staff. There is a possibility to get commercial support in case of interest. Prior to July 1, 2006 the package was named CDSware, then it was renamed to CDS Invenio, and now it is known simply as Invenio. The latest version of the Invenio is 3.0.1 [13]. The package was recently chosen to be the digital library software of some famous national universities around the world. [14] There are more than 60 registered institutions which are using the Invenio software. The number of new users is increasing on weekly basis.

Power of the Invenio is demonstrated in the CERN repository where the IR manages over 1,000,000 bibliographic records in high-energy physics since 2002, covering articles, books, journals, photos, videos, datasets, and more.

Invenio runs on Unix-like systems and requires Python/Flask web application server, MySQL, PostgreSQL or SQLite database server, Elasticsearch for information retrieval and Redis for caching. The IR has great multi-language support in this IR [15].

In the Invenio all resources (documents) are organized into collections [Figure 1]. Collections can be regular and/or virtual collections which are only important for the Invenio GUI. Collections can be customized in order to have different web interfaces, workflows and other features. Each collection can be customized for the general look and feel of its web pages. Linking rules are defined in order to implement relations between documents.

Basic entity in Invenio is the record [16] that contains metadata and may be associated with one or more documents (the digital content). Document can be stored in one or more revisions and a revision in one or more formats. Each record contains a unique identifier and can store articles, books, theses, photos, videos, museum objects and more. MARC is the standard metadata schema used in the Invenio, but other metadata sets can also be defined. Records can be submitted by an author or a librarian, through custom and fully configurable web interfaces. Workflows can be customized to create the proper steps for submission, review, conversion of documents, approval etc. Alternatively metadata and files can be ingested using customized conversion scripts, harvested from OAI-PMH compatible repositories or sent by e-mail. Also, this IR could serve as an OAI-PMH Data Provider and can exports records in MARCXML and BibTeX format and supports RSS feeds. Typically the MARC XML is natively used in the Invenio.

The Invenio is a modular framework of official and independent collaborative component packages. List of all are available official packages are described in documentation [13]. To be more precise all packages are located in two different GitHub organistations. Firstly, official packages are located in *inveniosoftware* collection [17]. The *inveniosoftware* is a collection of base packages, core packages, and additional feature packages that are maintained in a coherent, homogeneous way by the Invenio project team. Secondarily, community contributuons are situated in a collection of third-party packages extending [18] Invenio functionalities. They are maintained by contributing teams and may follow different practices. The packages may also incubate an experimental or unproven feature that can later mature into the main organisation. In the following, authors will address modules that can be used to create and exchange data from scientific research domain such as publications, books, monoghrah etc. The most important are those modules which can be involved in process of exchanging data between the Invenio and CERIF like systems.

The Invenio has numerous ways to import or export records' data. One of the most popular is *Invenio-oaiserver* which provides possibility to exchange information with other OAI-PMH compatible systems. CRIS UNS has opportunity to export and import PhD theses [19] with OAI-PMH protocol. Besides OAI-PMH, the Invenio has implemented *Dcxml* for generating XML representation in accordance with Dublin Core format [20] which is also supported format in CRIS UNS [10]. The Invenio has a support for Rest API with *Invenio-rest* and *Invenio-JSONSchemas*. This feature could be useful for the CRIS UNS due to the fact that it is REST compatible [21]. In the Invenio is implemented the package which could exchange data with famous repositories such as the OPENAIRE [22]. Data from this repository could be easily mapped to the CERIF format [23]. Evalution metrics of publication can be stored in the Invenio with the support of packages *DataCite* and *invenio-metrics*. This metrics are supported in CRIS UNS model [24]. Information about authors could be easily mapped to CERIF [25] because the Invenio has package for acquiring author's data based on unique authors' ids [26] by using *invenio-orcid*.

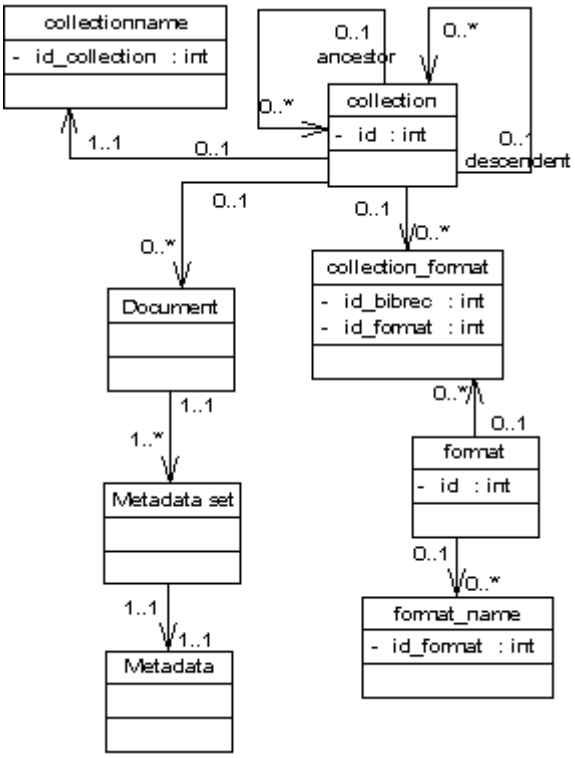

Figure 1 - Invenio data model

III. CERIF MODEL

CERIF is a standard that describes data model which can be used as a basis for an exchange of data from scientific-research domain. CERIF Standard describes the physical data model [27] and the exchange of XML messages between the CRIS systems [28]. The best feature of CERIF is that it can be expanded and adapted to different needs. In practice, CERIF is often mapped to other standards that also represent the data of scientific-research domain, for example CERIF/MARC21 mapping described in [29]. Authors of [30] recommend an extension of CERIF that incorporates a set of metadata required for storing theses and dissertations. Another example is [31] where authors argue how CERIF can be used as a basis for storage of bibliometric indicators.

Hereinafter we will present main entities of the CERIF data model version 1.5

- Base Entities - represent the core (basic) model entities. There are only three basic entities *cfPerson*, *cfOrganizationUnit* and *cfProject*.
- Result entities - A group of entities which includes results from scientific research like

publications, products and patents. Representatives of this group are: *cfResultPublication*, *cfResultProduct* and *cfResultPatent*.

- Infrastructure Entities - represent a set of infrastructure entities that are relevant for scientific research. The entities which belong in this group are: *cfFacility*, *cfEquipment* and *cfService*.
- 2nd Level Entities - Entities that further describe the Base Entities and Result Entities. E.g. *cfMedium* can be physical representation of some Result Entity.

Link Entities - are used to link entities from different groups. Typical entities of this group are: *cfOrganizationUnit_OrganizationUnit*, *cfOrganizationUnit_ResultPublication* and *cfResultPublication_DublinCore*. Link Entities allow for a generic classification mechanism to define their meaning, indicating the role for each entity instance in a relationship. Every Link

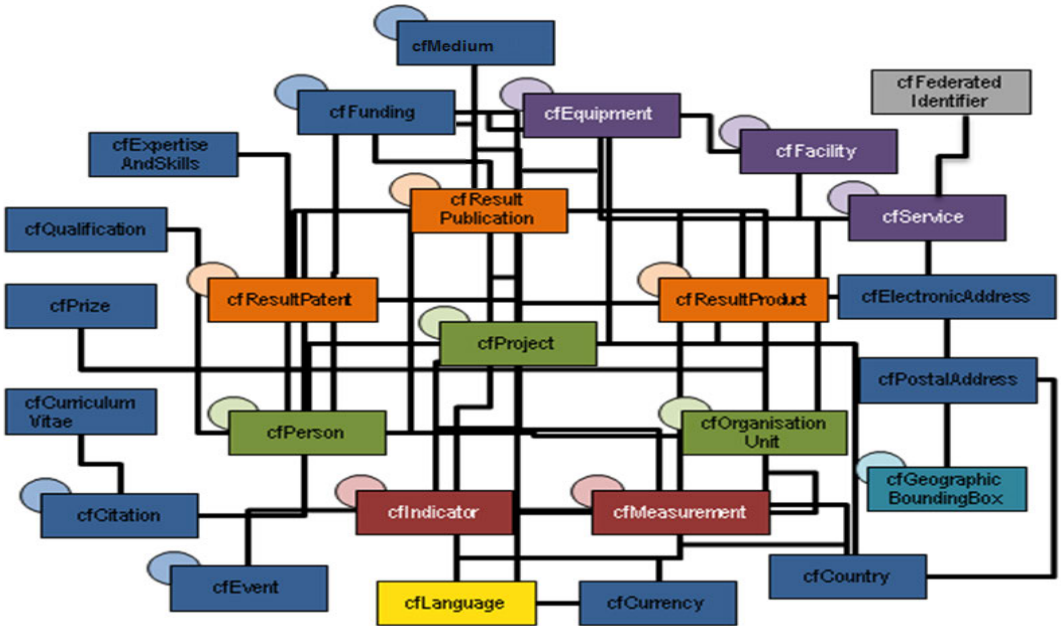

Figure 2 – CERIF model

entity is described with a role (*cfClass*, *cfClassScheme*), timeframe of relation (*cfStartDate*, *cfEndDate*), value (*cfFraction*) and identifiers of elements creating relation (e.g. *cfOrgUnit*, *cfResPublId*). The 'role' in link entities is not stored directly as attribute value, but as reference to Semantic layer.

- Multiple Language Entities - These entities provide multilingualism in CERIF for some entities.
- Semantic Layer Entities - Provide different kinds of semantics in CERIF model. The entities in this group are *cfClassificationScheme* and *cfClassification*. Those entities are used to describe classes and classification schemes for link and other entities. CERIF prescribes a controlled vocabulary to describe some of the classifications.
- Additional Entities - Currently in this group are classified entities that represent DC record.

Figure 2 [Figure 2] shows some of Base, Result, Link and Multiple Language Entities which are relevant for the mapping proposed in this paper.

IV. MAPPING SCHEME FOR INVENIO TO CERIF

The motivation of the authors for the mapping is arisen from the fact that they are part of the development team of the CRIS UNS system [3] which tends to be interoperable with the other systems and currently does not have ability to obtain data from the Invenio system. In the paper [32], a CERIF compatible research management system CRIS UNS is presented, which can be accessed at [33]. Currently, the system stores over 14,500 records of scientific publications (papers published in journals, conference proceedings, monographs, technical solutions and patents etc.). CRIS UNS system is under development since 2008 at the University of Novi Sad in the Republic of Serbia. Former development of that system covered implementation of the system for entering metadata about scientific research results [34]. Later phases in the development of CRIS UNS system included integration of various extensions that relay on CERIF model.

Proposed mapping scheme did not include any customization of the Invenio distribution since the most users use just the default installation.

Before mapping authors did specific analysis of the Invenio. Actually, the analysis of the software architecture, functionality, data model and implementation is conducted to determine: which data the system supports, in which manner and how the data is stored and is there any available module for exporting that data in some standardized formats. First the authors installed the IR and did an exploratory testing of software usage and functionality. After that, the comprehensive analysis of

TABLE I.
GREENSTONE DOCUMENTS METADATA FIELDS

Invenio Enitities	CERIF Entities	multiple	Cerif Link Enitity	Used classification
bib24X(tag="245 $a", value)	cfResPubl;cfResPublTitle (cfTtitle)	X		
bib10X(tag="100 $a", value)	cfPers; cfPersName (cfFirstName/cfLastName/cfOtherName)	X	cfPers_ResPubl	scheme: Person Output Contributions; Classificattion: Author
bib26X(tag="260 $b", value)	cfOrgUnit; cfOrgUnitName(cfName)		cfOrgUnit_ResPubl	scheme: Organisation Output Roles; Classificattion: Publisher
bib26X(tag="260 $a", value)	cfPAddr (cfCityTown)		cfOrgUnit_PAdrr	scheme: Organisation Contact Details; Classificattion: Postal Address
bib26X(tag="260 $c", value)	cfResultPubl(cfResPublDate)			
bibdoc (creation_date, octype, docname); bibrec_bibdoc	cfMedium cfMediumCreationDate, cfMimeType);cfMediumTitle(cfTitle)	X	cfResPubl_Medium	scheme: Organisation Contact Details; Classificattion: Postal Address

software architecture and its components was done. Functionality of the relevant modules, their interconnections, and the database is inspected more closely based on software documentation [13] and source code. Moreover, the authors identified available modules that support data exchange between the Invenio and the other repositories with special emphasis on the CRIS systems.

Data in the Invenio are stored in the records which are in accordance to the MARC 21 standard. The system provides data export in the MARCXML [35] that could be beneficial for the systems that are the MARC 21 compatible [34]. Also, the Invenio has a service (Invenio-oaiserver) for exchanging data via the OAI-PMH protocol. Despite the fact that the CRIS UNS and some systems have support for mentioned standards [19], vast majority of the IRs and CERIF compatible systems do not implement the MARC 21 and the OAIPMH. Thus, the Invenio data should be represented in accordance to well-known CERIF standard for representing scientific research data.

All the Invenio's records are represented with *bibrec* entity [Figure 3]. The *bibrec* is the primary element for representing the MARC 21 data. Concrete records' values are actually stored in the auxiliary entities from the *bib00X* to the *bib99X* which are keeping the real MARC 21 data, for instance, value of MARC 21 representation "245$a" as a title of a publication will be stored in *bib24X* table. Link between *bibrec* and some *bibXYX* is achieved with entities *bibrec_bibXYX*. The Invenio records are hierarchically organized in the collections. There are primary (regular) and secondary collections (virtual) [13] where the primary collections are basic organizational types (Books, Theses, Articles, Preprints, Reports, Pictures …) of how the records are grouped together in collections, while the secondary are used mostly for improving web UI and search results. The Collections are not directly connected [Figure 3] to the record as it is expected. This link between particular record and the collection is established through MARC 21 field 980 and subfield $a for the bibliographic record. So, all records from one collection could be retrieved from entity bib98X where field tag is "980 $a" and field value holds the name of the collection e.g. "Theses". The Table 1 presents a proposal for mapping specific the Invenio record types to the adequate CERIF entities. The first column is related to the indentified *bibrec* types from the default installation of the Invenio. The second column is reserved for the appropriate CERIF entity which in this case is the *cfResPubl*. The third column The fourth and the fifth columns are used to provide different kinds of semantics in the CERIF model for the *cfResPubl* by using powerful semantic layer of CERIF model. The CERIF predefined classifications and classifications schemes are used to represent the primary collections of the Invenio. In the following, it is presented how to map the record which is a member of the book collection where a part of the relevant mapping information is shown in the Table 1. This record is represented via appropriate MARC fields (entities from the Invenio model). The Book Title is described with the Invenio entity *bib24X* which field tag has value "245 $a" and field value with actual name of the book e.g. "Game of thrones". The title value will be mapped to the CERIF entity *cfPublTitle*, precisely in its attribute *cfTitle*. The author of the book is stored in the Invenio entity *bib10X* which field tag has value "100 $a" and field value with actual name of the author e.g. "Martin, George R.R". The author is mapped to the CERIF entity *cfPers* and its name is stored in the fields cfFirstName/cfLastName/cfOtherName of the enitity *cfPersName*. Entity *cfPers* is linked to the *cfResultPubl* via the entity *cfPers_ResPubl*. The *cfPers_ResPubl* need to be classified with the CERIF semantic layer by using the scheme Person Output Contributions and the Classification Author. This semantic organization is very useful when it is important to distinguish different person roles such as authors, editors and reviewers. Every book record usually has the book publisher that is represented with the Invenio entity *bib26X* which field *tag* has value "260 $b" and field value with actual name of the publisher e.g. "Harper Voyager". The publisher data is mapped to the CERIF entity *cfOrgUnit* and its name will be defined in the *cfOrgUnitName* where the concrete value will be stored in its attribute *cfName*. Entity *cfOrgUnit* is linked to

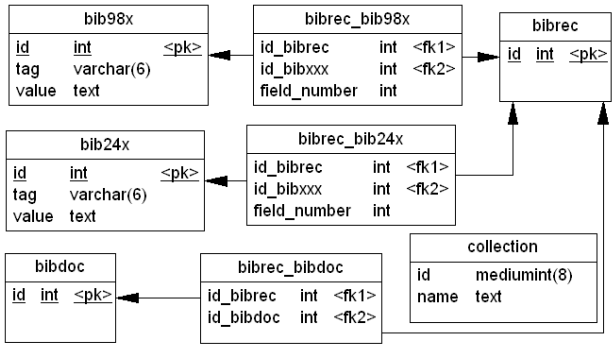

Figure 3

the *cfResultPubl* via entity *cfOrgUnit_ResPubl*. The *cfOrgUnit_ResPubl* need to be classified with CERIF semantic layer by using scheme Organisation Output Roles and Classification Publisher.

The place of the publication is represented with the Invenio entity *bib26X* which field tag has value "260 $a" and field value with real name of the place e.g. "London". The data about the location will be mapped to the CERIF entity *cfPAddr* and the concrete value will be stored in its attribute *cfCityTown*. The Entity *cfPAddr* is linked to the *cfOrgUnit* via entity *cfOrgUnit_PAdrr*, providing the information about the place of the publication. The year of the Publication is represented with the Invenio entity *bib26X* which field tag has value "260 $c" and field value with actual year of publishing e.g. "1998". Publication year will be stored to the CERIF entity *cfResultPubl*, precisely in the attribute *cfResPublDate*. Some of the Invenio records have attached digital resources such as doc, xls, pdf, image files, etc. The information for the digital resources in the Invenio is stored in entity *bibdoc* that is linked with the Invenio record via entity *bibrec_bibdoc*. All the digital resources from the Invenio are represented as the instances of the CERIF *cfMedium* entity, in accordance to EuroCRIS suggestion [36], whereas the value of the *bibrec_bibdoc* is mapped to the CERIF *cfResPubl_Medium* entity. The attributes *creation_date* and *doctype* from *bibdoc* entity are directly mapped to the attributes *cfMediumCreationDate* and *cfMimeType* of the *cfMedium*. The attribute *docname* of the *bibdoc* is mapped to the *cfTitle* attribute of the entity *cfMediumTitle* which has a link with the *cfMedium*. The *cfResPubl_Medium* needs to be classified with CERIF semantic layer by using scheme Organisation Contact Details and Classificattion Postal Address.

V. CONCLUSION

The importance of institutional repositories and CRIS systems for scientific research data is enormous. Making data accessible between these systems is unavoidable. Therefore, this paper presents a mapping scheme for the Invenio data to CERIF model where all the Invenio data is available to CERIF like systems (CRISs, IRs that support CERIF, etc.).

The main contributions of this research are:

• Proposal for mapping data from Invenio repository to the current 1.5 CERIF model

• Potential using of Import/Export plug-in for making full interoperability between these two systems.

Future work will be directed towards mapping the data from other IRs such as Fedora and SobekCM in the CERIF format.

ACKNOWLEDGMENT

Results presented in this paper are part of the research conducted within the Grant No. III-47003, Ministry of Science and Technological Development of the Republic of Serbia.

REFERENCES

[1] N. F. Foster and S. Gibbons, "Understanding faculty to improve content recruitment for institutional repositories," -Lib Mag., vol. 11, no. 1, pp. 1–12, 2005.

[2] Á. Rocha, A. M. Correia, S. Costanzo, and L. P. Reis, New contributions in information systems and technologies. Volume 1 Volume 1. 2015.

[3] "Invenio." [Online]. Available: http://invenio-software.org/. [Accessed: 20-Dec-2014].

[4] "Welcome :: Greenstone Digital Library Software." [Online]. Available: http://www.greenstone.org/. [Accessed: 20-Dec-2014].

[5] "EPrints - Digital Repository Software." [Online]. Available: http://www.eprints.org/. [Accessed: 20-Dec-2014].

[6] "DSpace | DSpace is a turnkey institutional repository application." [Online]. Available: http://www.dspace.org/. [Accessed: 20-Dec-2014].

[7] "Fedora Repository | Fedora is a general-purpose, open-source digital object repository system." [Online]. Available: http://fedora-commons.org/. [Accessed: 20-Dec-2014].

[8] "SobekCM : Digital Content Management System." [Online]. Available: http://sobekrepository.org/.

[9] "Common European Research Information Format | CERIF," 2000. [Online]. Available: http://www.eurocris.org/. [Accessed: 18-Jan-2014].

[10] V. Penca and S. Nikolić, "Scheme for mapping Published Research Results from Dspace to Cerif Format," in 2. International Conference on Information Society Technology and Management, 2012, pp. 170–175.

[11] Valentin Penca, Siniša Nikolić, and Dragan Ivanović, "Scheme for mapping scientific research data from EPrints to CERIF format," 2015.

[12] V. Penca, S. Nikolić, and D. Ivanović, "Mapping scheme from Greenstone to CERIF format," in Proceedings of the 6th International Conference on Information Society and Technology (ICIST 2016), Kopaonik mountain resort, Republic of Serbia, 2016.

[13] CERN, "invenio Documentation." 07-Dec-2016.

[14] "Current Invenio Users." [Online]. Available: http://inveniosoftware.org/showcase.

[15] S. R. Lihitkar and R. S. Lihitkar, "Open Source Software for Developing Digital Library: Comparative Study," DESIDOC J. Libr. Inf. Technol., vol. 32, no. 5, pp. 393–400, Sep. 2012.

[16] G. Pyrounakis, M. Nikolaidou, and M. Hatzopoulos, "Building Digital Collections Using Open Source Digital Repository Software:: A Comparative Study," Int. J. Digit. Libr. Syst., vol. 4, no. 1, pp. 10–24, 2014.

[17] "Invenio Software Official Libraries." [Online]. Available: https://github.com/inveniosoftware.

[18] "Invenio Software Contributor Libraries." [Online]. Available: https://github.com/inveniosoftware-contrib.

[19] L. Ivanović, D. Ivanović, and D. Surla, "Integration of a Research Management System and an OAI-PMH Compatible ETDs Repository at the University of Novi Sad," Libr. Resources Tech. Serv., vol. 56, no. 2, pp. 104–112, 2012.

[20] "Dublin Core® Metadata Initiative (DCMI) Format," 2014. [Online]. Available: http://dublincore.org/. [Accessed: 11-Sep-2014].

[21] V. Penca, S. Nikolić, and D. Ivanović, "SRU/W service for CRIS UNS system," in 4. International Conference on Information Science and Technology (ICIST), Kopaonik, 2014, pp. 108–114.

[22] "OpenAIRE." [Online]. Available: https://www.openaire.eu/.

[23] "OpenAIRE Guidelines for CRIS Managers." [Online]. Available: http://eurocris.org/openaire-gl.

[24] R. Gartner, M. Cox, and K. Jeffery, "A CERIF-based schema for recording research impact," Electron. Libr., vol. 31, no. 4, pp. 465–482, 2013.

[25] Demeranville, Tom, "ORCID and CERIF," in euroCRIS Strategic Membership Meeting Autumn 2015, Barcelona, 2015.

[26] "ORCID Connecting Research and Researchers." [Online]. Available: https://orcid.org/.

[27] B. Jörg et al., CERIF 1.3 Full Data Model (FDM) Introduction and Specification. 2012.

[28] J. Dvořák and B. Jörg, "CERIF 1.5 XML - Data Exchange Format Specification," 2013, p. 16.

[29] D. Ivanović, D. Surla, and Z. Konjović, "CERIF compatible data model based on MARC 21 format," Electron. Libr., vol. 29, pp. 52–70, 2011.

[30] L. Ivanovic, D. Ivanovic, and D. Surla, "A data model of theses and dissertations compatible with CERIF, Dublin Core and EDT-MS," Online Inf. Rev., vol. 36, no. 4, pp. 548–567, 2012.

[31] S. Nikolić, V. Penca, and D. Ivanović, "STORING OF BIBLIOMETRIC INDICATORS IN CERIF DATA MODEL," Kopaonik mountain resort, Republic of Serbia, 2013.

[32] D. Surla, D. Ivanovic, and Z. Konjovic, "Development of the software system CRIS UNS," in Proceedings of the 11th International Symposium on Intelligent Systems and Informatics (SISY), Subotica, 2013, pp. 111–116.

[33] "Current Research Information System of University of Novi Sad." [Online]. Available: http://www.cris.uns.ac.rs/. [Accessed: 18-Jan-2014].

[34] D. Ivanović, G. Milosavljević, B. Milosavljević, and D. Surla, "A CERIF-compatible research management system based on the MARC 21 format," Program Electron. Libr. Inf. Syst., vol. 44, no. 3, pp. 229–251, 2010.

[35] "MARCXML: The MARC 21 XML Schema." [Online]. Available: http://www.loc.gov/standards/marcxml/schema/MARC21slim.xsd. [Accessed: 18-Jan-2014].

[36] Houssos, Nikos, Jörg, Brigitte, and Matthews, Brian, "A multi-level metadata approach for a Public Sector Information data infrastructure," presented at the CRIS2012 Conference, 2012.