SEARCH OF CATALOGUES OF THESES AND DISSERTATIONS

Lidija Ivanović[1]

Abstract. The aim of this paper is to review the existing catalogues of dissertations available through Internet as well as standards on which those catalogues are made. Various systems that store dissertations metadata were analyzed. The analyses found out that these systems are based on a number of standards, protocols and formats that enable interoperability of these systems: CERIF standard, AOI-PMH protocol, Dublin Core format, MARC 21 format, ETD-MS format. Based on this analysis within the CRIS UNS system is implemented a module for search of PhD dissertations in such a way that the dissertations metadata are in accordance with previously mentioned standards, protocols and formats.

AMS Mathematics Subject Classification (2010): 68U35

Key words and phrases: theses and dissertations, digital libraries, NDLTD, DART-Europe, CRIS UNS

1. Introduction

Nowadays, an important part of scientific knowledge is available through published research results in digital format along with associated metadata in a variety of applications such as digital libraries, institutional repositories, library information systems, research management systems. A catalogue of scientific results (in most of these systems) can be browsed and searched using the Internet. The importance of public access to research results published in digital form for further development of science is the main topic of the papers [17, 5, 2, 1, 14, 15, 16, 4]. This paper provides an overview of existing systems that store metadata about theses and dissertations in the digital form. The various types of systems that have metadata about dissertations were analyzed. The analyses found out that these systems are based on a number of standards and protocols that allow interoperability: CERIF standard (*Common European Research Information Format* - `http://www.eurocris.org/Index.php?page=CERIFintroduction&t=1`), AOI-PMH protocol (*Open Archives Initiative Protocol for Metadata Harvesting* - `http://www.openarchives.org/pmh/`). This paper describes these systems as well as relevant standards. There is a need for the integration of the analyzed systems in order to increase the availability of dissertations without duplicate entry of metadata on different platforms. This integration can be done by forming a corresponding set of metadata that describe a dissertation. Such integration is presented in the example of the development of the digital library of dissertations defended at the University of Novi Sad.

---
[1]University of Novi Sad, Faculty of Education in Sombor,
e-mail: lidija.ivanovic@pef.uns.ac.rs

2.   Digital libraries of theses and dissertations

Digital libraries of theses and dissertations are designed to allow rapid search and retrieval of theses and dissertations in the digital form. In recent years, there is a trend of integration of digital libraries in networks of digital libraries. Their basic idea is that each research institution, that is a member of the network, has its own repository of theses and dissertations and allows retrieval of metadata about its theses and dissertations by an adopted protocol (e.g. OAI-PMH). It enables creation of a unified catalogue of theses and dissertations from all over the network, which increases the availability of theses and dissertations. This section describes the public catalogue of dissertations of the Centre for Scientific Libraries (CRL Online Catalog). It also describes the two networks of digital libraries of theses and dissertations: World network (NDLTD) and European network (DART-Europe) of digital libraries of theses and dissertations.

2.1.   CRL Online Catalog

At the end of the Second World War, the ten universities (*Midwestern research universities*) made a common repository for materials called *Midwest Inter-Library Corporation* (MILC) [19]. Most libraries during the period 1950-1960 were overcrowded by books. Therefore this common repository MILC was created. MILC in the 1970s changed its name to the CRL (*The Center of Research Libraries*). CRL is an international consortium of universities, colleges and independent research libraries available on this web site `www.crl.edu`. It collects and stores dissertations, monographs, journals, documents, archives, and other traditional and digital resources from a global network of sources. At the end of 2012 CRL had more than 250 members. In addition, the material CRL receives from the United States and Canada are not the only one, it also receives materials from five regions: Africa, the Middle East, Southeast Asia, South Asia and Latin America. CRL Online Catalog (`http://catalog.crl.edu/search~S4`) enables search of catalogue of dissertations. The total number of PhD dissertations which are not from the U.S.A. and Canada, at the end of 2012 was over 800,000. CRL enables search by *keyword, title, author, subject*, as well as search by name of the *country* and the *institution* where the dissertation is defended. If user selects a country from the drop-down list, the system displays all institutions from the selected country whose dissertations exist in the system. Basic search is shown in **Figure 1**.

Moreover, there is advanced search of dissertations which includes search by *author, title, subject* with possibility to use the followings operands AND, AND NOT, OR, NEAR, BEFORE, AFTER (**Figure 2**). Metadata of searching results are represented in MARC 21 format.

2.2.   NDLTD

NDLTD (*Network Digital Library of Thesis and Dissertations* - `http://www.ndltd.org/`) is an international organization whose aim is to create a world-wide network of theses and dissertations in the digital form. This organization promotes open access to scientific knowledge that exists in the theses and dissertations of researchers from all over the world. At the end of 2012, NDLTD had

Figure 1: CRL basic search

Figure 2: CRL advanced search

more than one million theses and dissertations in the digital form and hundreds of institutions and individuals who are members of the network. The network has a central node and hundreds of peripheral nodes which contain theses and dissertations of universities that are members of the network. It is necessary to allow access to the repository of theses and dissertations via the OAI-PMH protocol for becoming a member of the NDLTD. OAI-PMH is the protocol for exchanging metadata which can be represented in various metadata formats such as: Dublin Core, MARC 21, ETD-MS, etc. The NDLTD network demands the use of the ETD-MS format (*Electronic Theses and Dissertations Metadata Set* - `http://www.ndltd.org/standards/metadata/etd-ms-v1.00-rev2.html`) for representing theses metadata. So, the central node of the NDLTD network can import metadata about theses represented using the ETD-MS format via the OAI-PMH protocol. NDLTD does not require from members of the network that metadata of theses and dissertations are publicly available, but it is recommended. The OAI compatibility of digital library of theses and dissertations can be checked before turning the institution into the network member using applications available at `http://re.cs.uct.ac.za/`. Authors whose institutions do not belong to the network can enter their theses and dissertations metadata through a web application of the central node. The server of the central node

periodically updates metadata about theses and dissertations whose digital
content is found in the peripheral nodes. In addition to title, author's name
and other basic metadata, the ETD-MS format prescribes the metadata *identi-
fier* which holds URL linking to the thesis or dissertation page on a peripheral
node. The metadata list in this format is shown in **Table 1**.

Metadata about thesis or dissertation	Element ETD-MS format
author	dc.creator
advisor	dc.contributor
chair	dc.contributor
committee member	dc.contributor
title	dc.title
alternative	title dc.title.alternative
keywords	dc.subject
abstract	dc.description.abstract
note	dc.description.note
language	dc.language
publisher	dc.publisher
publication date	dc.date
record type	dc.type
content format	dc.format
URL	dc.identifier
access rights	dc.rights
name of author's degree after defense	thesis.degree.name
level of education	thesis.degree.level
scientific field	thesis.degree.discipline
institution	thesis.degree.grantor

Table 1: ETD-MS format - metadata about theses and dissertations

The NDLTD unique catalogue can be browsed and searched by two specially
developed web applications: *Scirus ETD Search* www.ndltd.org/serviceprov
iders/scirus-etd-search and *VTLS Visualizer* http://thumper.vtls.com:
6090/. **Scirus ETD Search** is a comprehensive search tool (**Figure 3**). It
offers *basic* and *advanced search*. Selection of a retrieved thesis or dissertation
can: immediately open the pdf format of the thesis/dissertation; open a new
window with metadata and URL (*access link*) to the thesis/dissertation; open
the page within repository of institution which the thesis/dissertation belongs
to. NDLTD allows that members of the network can define the access rights
and privacy of data and that is the reason for various situations happening after
selection of a retrieved thesis or dissertation. Some metadata are immediately
visible and some are not.

VTLS Visualizer is a web application that offers sophisticated basic and
advanced search. It allows searching of parts of NDLTD catalogue by selecting
the appropriate set or the entire NDLTD network by *language* on which the
theses are written (English, German, French, etc.), by *format* of digital form of
theses (pdf, html, doc, etc.), by *year of publication*, by *institution*, by *scientific*

Figure 3: Scirus ETD tool for search of theses and disseratations

field. After retrieving results, the user can choose a thesis or dissertation which is displayed in basic form that contains link *View Source Record*. The link opens a page with URL to the thesis or dissertation and the option *Show full item record* that opens metadata represented in the Dublin Core format. **Figure 4** shows an *advanced search* of the catalogue.

Figure 4: VTLS Visualizer - advanced search

2.3. DART-Europe

DART-Europe (www.dart-europe.eu) promotes open access to scientific knowledge that exists in theses and dissertations of researchers from Europe. Members of this group can become institutions, consortia and national repositories. **DART-Europe** collects metadata using the OAI-PMH protocol in the

Metadata about thesis or dissertation	Element of Dubline Core format
author	dc.creator
title	dc.title
keywords	dc.subject
abstract	dc.description
language	dc.language
publisher	dc.publisher
publication date	dc.date
record type	dc.type
content format	dc.format
URL	dc.identifier
access rights	dc.rights

Table 2: Dublin Core format - metadata about theses and dissertations

Dublin Core format. Metadata in the Dublin Core format are shown in **Table 2**.

At the end of 2012 this network contained 362,457 theses from 523 Universities from 27 European countries. This catalogue of theses and dissertations can be searched through basic and advanced search. Advanced search allows search by *country*, by *year of publication*, by *author name*, by *repository* or *group of repositories* which are members of the network, by *university* and by *language*. Results of searching can be sorted (in ascending / descending order) by *title*, *author name*, *year of publication*, *university name* and *repository name*. Selection of a retrieved thesis or dissertation opens a page with metadata represented in the Dublin Core format and an URL downloading thesis or dissertation in digital format (pdf).

3.   Different storage systems of theses and dissertations

In addition to digital libraries of theses and dissertations, there are other systems that contain metadata about theses and dissertations. The most famous among them are:

- Institutional repositories

- Library information systems

- Research management systems

An **institutional repository** is a software system for collecting, preserving and disseminating digital contents that present published results of researchers employed at a scientific institution. These digital contents can be published as journal articles, papers presented at a scientific conference, master's theses, doctoral dissertations, and so on. Besides the possibility to download digital contents, most of these systems allow downloading of metadata about digital contents according to the protocol OAI-PMH in the Dublin Core format. There are many institutional repositories and the most used platforms for creating

institutional repositories are Eprints (`www.eprints.org/`) and DSpace (`www.dsp ace.org/`). There is a trend to create networks of institutional repositories in order to create virtual repositories. An example of such a network is DRIVER (`www.driver-repository.eu/`). DRIVER is a European open access network of institutional repositories. It is created with the aim to increase the availability of knowledge. Member of the network can become an institutional repository that is OAI compatible and stores metadata in Dublin Core format.

Library information systems are implemented in accordance with the library standards (the most commonly used standard is MARC 21 format) and intended for cataloguing library materials. Library information systems contain a lot of catalogued published scientific-research results, among which are theses and dissertations. There are many library information systems and some of them are: ALEPH 500, Atriuum, COBISS, BISIS, etc.

Research management systems are intended to store all data relevant for scientific-research domain: research projects, scientific research institutions, researchers, equipments necessary for research, scientific conferences, scientific research published results. The central part of these systems is storing of the scientific research results, among which are a lot of theses and dissertations. The European Union encourages the development of research management systems on the standardized CERIF data model with aim to enable exchanging of data between these systems. Some of research management systems are: CRIStin, HunCris, SICRIS, ISTWorld, CRIS UNS, etc.

4. Integration of systems storing theses and dissertations

Integration of the systems that store theses and dissertations are intended to avoid or reduce input of the same metadata in different systems. One such system is the NARCIS (`www.narcis.nl`). NARCIS is the national portal of Netherland, designed for searching information about the researchers and their work. It provides access to scientific information from all research institutions and repositories of all Netherland universities. Collecting data within NARCIS is described in the paper [3]. At the end of 2012 NARCIS contained 46,791 PhD dissertations, 727,981 publications, 2914 organizations. Search of dissertations is possible by *year*, *institution* and *accessibility* (open/close accessibility). The paper [6] provides a table which shows that CRIStin contains records imported from other systems. In addition to basic and advanced search, the CRIStin information system provides searching via Norwegian Open Research Archives (NORA - `www.ub.uio.no/nora/search.html`). NORA is a service that allows search of all institutional repositories and open access journals from Norway. The paper [20] describes the cooperation between research management systems and digital libraries of Pretoria University. Similar cooperation between research management systems and digital libraries exist at *Glasgow University*, *Southampton University* and *Kingston University*. The aim of the group CRIS-IR (`www.eurocris.org/Index.php?page=CRIS-IR_workplan&t=1`) is finding an optimal solution that enables interoperability of research management systems on the one hand and institutional repositories on the other, taking into account all relevant aspects. The aim of CRIS-OAR Interoperability

Project (`www.knowledge-exchange.info/Default.aspx?ID=340`) is to increase interoperability between research management systems and open access digital repositories by defining exchange metadata format.

5. Search of dissertations defended at the University of Novi Sad

In order to obtain a system that is compatible with all systems that contain theses and dissertations the creation of a rich data model is a necessity. That data model should contain all the metadata prescribed by standards on which those systems are built. One such model is the CERIF compatible data model based on the MARC 21 format described in the paper [8]. In this compatible model data defined by CERIF data model about publications, patents, products, people, projects, events and organizations are mapped to the MARC 21 format (bibliographic or authority format). This compatible model includes attributes of all CERIF entities and preserves the existing relationship between the entities of the CERIF data model. It also enables the entry of multilingual information prescribed by the CERIF data model. MARC 21 format is rich with metadata and it allows a more detailed description of the entity of research management systems than the CERIF data model. The format can store all metadata required by Dublin Core and EDT-MS format [10]. Information systems based on the CERIF compatible data model can exchange data with research management systems via XML documents prescribed by CERIF standard, can exchange data in MARC format with library information systems, as well as in Dublin Core and ETD-MS format with networks of repositories [11]. On the basis of this model, the digital repository of dissertations of the University of Novi Sad is built. This repository is part of the research management system of the University of Novi Sad (CRIS UNS - `http://cris.uns.ac.rs/index.jsf`). The papers [7, 18] describe the implementation of the system CRIS UNS. Automatic extraction of metadata from published results stored in CRIS UNS system is described in [13]. The published results from this system are publicly available via Internet and their metadata are entered by the authors who do not need to know CERIF standard and MARC 21 format. Moreover, besides international requirements prescribed by CERIF standard, the system meets the requirements of the *Ministry of Education, Science and Technological Development of the Republic of Serbia* within the results evaluation domain. Therefore, the data model of the system is expanded with necessary entities as it is described in the papers [9, 12].

Within the system CRIS UNS is implemented a special module for searching dissertations that were defended at the University of Novi Sad (`http://cris.uns.ac.rs/searchDissertations.jsf`). At the end of 2012 the number of defended PhD dissertations was around 4000. Dissertations can be searched in three different modes:

1. Search of dissertations by forming boolean query using user interface elements

2. Search by authors and board members

3. Search by Lucene query language

The most commonly used is the first mode, which is present in **Figure 5**. This mode enables search of dissertations by the followings fields: **Full Text; Title, Abstract, Keywords; Title; Abstract; Keywords; Author; Advisor; Committee chairman; Committee member**. The operator between these field can be **AND, OR, AND NOT**. Results can be filtered by publication years, by institutions, and by gained titles. The results of search are displayed as a reference in Harvard style. Moreover, there are options for obtaining additional information that include metadata about dissertations in MARC 21, Dublin Core and ETD-MS format.
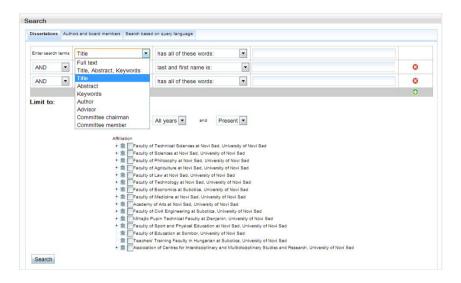

Figure 5: Search of dissertations defended at the University of Novi Sad

6.   Conclusion

This paper describes the systems that store metadata about theses and dissertations, as well as theses and dissertations in digital form. Obviously, there are a variety of systems and standards on which these systems are built. Furthermore, the paper provides a special emphasis on the ability of exchanging data between these systems that could increase the availability of theses and dissertations. Therefore, the increasing of availability of knowledge contained in the theses and dissertations should encourage further development of science. The analysis presented in this paper provided a list of functional requirements, as well as a list of standards that a publicly available catalogue of dissertations should support. This list includes the needs of dissertations catalogue's users related to browsing and searching catalogue, as well as needs for exchanging the metadata with other systems that contain theses and dissertations. The system implementation (shown in **Figure 5**) of the catalogue of dissertations

of the University of Novi Sad is described. It is implemented as a module within the CRIS UNS system, and it is compatible with the systems and standards analyzed in this paper. Future steps include the possibility of becoming a member of the NDLTD and DART-Europe networks of digital libraries of dissertations, as well as becoming a member of the DRIVER Europe network of institutional repositories. This will be possible because during the repository design standards, protocols and systems described in this paper were taken into account.

References

[1] Anderson, K., Sack, J., Krauss, L. and O'Keefe, L., Publishing online-only peer-reviewed biomedical literature: Three years of citation, author perception, and usage experience, Journal of Electronic Publishing, (2001), Vol. 6, No. 3, available at: `http://quod.lib.umich.edu/cgi/t/text/text-idx?c=jep;view=text;rgn=main;idno=3336451.0006.303` (accessed 30 December 2012).

[2] Antelman, K., Do open-access articles have a greater research impact?, College & Research Libraries, (2004), Vol. 65, pp. 372-382.

[3] Dijk E., NARCIS - more than a gateway to Open Accesspublications in The Netherlands, Open Access Infractructures: the Future of Scientific Communication, Athens-Greece, (2008), available at: `http://openaccess.gr/dotAsset/13716.pdf` (accessed 30 December 2012).

[4] Eysenbach, G., Citation advantage of open access articles, PLoS Biology, (2005), Vol. 4, No. 5, pp. 692-698.

[5] Harnad, S., Brody, T., Comparing the impact of Open Access (OA) vs. non-OA articles in the same journals, D-Lib Magazine, (2005), Vol. 10, No. 6, available at: `http://www.dlib.org/dlib/june04/harnad/06harnad.html` (accessed 30 December 2012).

[6] Hill, A., The Names Project for Author ID in the UK: how should national projects coexist with ORCID?, (2012), available at: `http://openaccess.uoc.edu/webapps/o2/bitstream/10609/16342/1/Names%20GrandIR%20meeting%206%20Sep%202012.pdf` (accessed 30 December 2012).

[7] Ivanović, D., Milosavljević, G., Milosavljević, B., Surla, D., A CERIF-compatible research management system based on the MARC 21 format, Program: Electronic libarary and information systems, (2012), Vol. 44, No. 1, pp. 229-251, DOI: 10.1108/00330331011064249.

[8] Ivanović, D., Surla, D., Konjović, Z., CERIF compatible data model based on MARC 21 format, The Electronic Library, (2011a), Vol. 29, No. 1, pp. 52-70, DOI: 10.1108/02640471111111433

[9] Ivanović, D., Surla, D., Racković, M., A CERIF data model extension for evaluation and quantitative expression of scientific research results, Scientometrics, (2011b), Vol. 86, No. 1, pp. 155-172, DOI 10.1007/s11192-010-0228-2.

[10] Ivanović, L., Ivanović, D., Surla, D., A data model of theses and dissertations compatible with CERIF, Dublin Core and EDT-MS, Online Information Review, (2012a), Vol. 36, No. 4, pp. 568-586, DOI 10.1108/14684521211254068.

[11] Ivanović, L., Ivanović, D., Surla, D., Integration of a Research Management System and an OAI-PMH Compatible ETDs Repository at the University of

Novi Sad, Republic of Serbia, Library Resources & Technical Services, (2012b), Vol. 56, No. 2, pp. 104-112.

[12] Ivanović, D., Surla, D., Racković, M., Journal evaluation based on bibliometric indicators and the CERIF data model, Computer Science and Information Systems, (2012c), Vol. 9, No. 2, pp. 791-811, DOI 10.2298/CSIS110801009I.

[13] Kovacević, A., Ivanović, D., Milosavljević, B., Konjović, Z., Surla, D., Automatic extraction of metadata from scientific publications for CRIS systems, Program: electronic library and information systems, (2011), Vol. 45, No. 4, pp.376-396, DOI: 10.1108/00330331111182094.

[14] Kürtz, M. J., Eichhorn, G., Accomazzi, A., Grant, C., Demleitner, M., Henneken, E., et al., The effect of use and access on citations, Information Processing and Management, (2005a), Vol. 41, No. 6, pp. 1395-1402.

[15] Kürtz, M. J., Eichhorn, G., Accomazzi, A., Grant, C., Demleitner, M., Murray, S. S., Worldwide use and impact of the NASA astrophysics data system digital library, Journal of the American Society for Information Science and Technology, (2005b), Vol. 56, No. 1, pp. 36-45.

[16] Kürtz, M. J., Eichhorn, G., Accomazzi, A., Grant, C., Demleitner, M., Murray, S. S., et al., The bibliometric properties of article readership information, Journal of the American Society for Information Science and Technology, (2005c), Vol. 56, No. 2, pp. 111-128.

[17] Lawrence, S., Free online availability substantially increases a paper's impact, Nature, (2001), Vol. 411, paper 521.

[18] Milosavljević, G., Ivanović, D., Surla, D., Milosavljević, B., Automated construction of the user interface for a CERIF-compliant research management system, The Electronic Library, (2011), Vol. 29, No. 5, pp. 565-588, DOI: 10.1108/02640471111177035.

[19] Mullins, J.L., The changing definition and role of collections and services in the University Research Library, Indiana Libraries, (2012), Vol. 31, No. 1, pp. 18-24, available at: `http://journals.iupui.edu/index.php/IndianaLibraries/article/viewFile/2061/1940` (accessed 30 December 2012).

[20] Olivier, E., Open Scholarship & research reporting in tandem: creating more value, The African Digital Scholarship & Curation Conference, May 12-14, (2009), available at: `http://repository.up.ac.za/bitstream/handle/2263/10109/Olivier_Open%282009%29.pdf?sequence=1` (accessed 30 December 2012).