Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science00 (2014) 000–000

Procedia
Computer Science

www.elsevier.com/locate/procedia

CRIS 2014

Electronic theses and dissertations in CRIS

Joachim Schöpfel[1]*, Danica Zendulkova[2], Omid Fatemi[3]

[1]*University of Lille 3, Villeneuve d'Ascq, France*
[2]*Slovak Centre of Scientific and Technical Information, Bratislava, Slovakia*
[3]*University of Tehran and Iranian Institute for Information Science & Technology, Tehran, Iran*

Abstract

Electronic theses and dissertations (ETD) represent a significant part of academic publications. They contain valuable information about academic research, in particular on research projects, institutions and experts. This information can be useful for the management of expertise and skills of persons and organisations in the current research information systems (CRIS). The paper provides an overview on projects and initiatives linking ETD and CRIS infrastructures, with empirical insight from existing systems in Slovakia, Iran and France. The paper reviews also the way the Common European Research Information Format (CERIF) integrates the specific information related to ETD (results, links, second level elements, semantics…). The discussion puts the focus on metadata, interoperability and complementary material (data). The findings allow for the framing of some recommendations on the integration of ETD in CRIS.
© 2014 The Authors. Published by Elsevier B.V.
Peer-review under responsibility of euroCRIS.

Keywords: Electronic theses and dissertations; current research information systems; metadata;CERIF

1. Introduction

Electronic theses and dissertations (ETD) represent a significant part of academic publications. Especially doctoral or PhD theses contain the results of at least three years of individual scientific work, accomplished in a laboratory, a research team or an institute, school or company. They are produced by universities, as part of academic grey literature, considered as “library material” and disseminated in limited numbers, with a specific legal status (Juznic 2010). They cover all scientific disciplines and represent up to 10% of national scientific output.

* Corresponding author. Email: joachim.schopfel@univ-lille3.fr

1877-0509© 2014 The Authors. Published by Elsevier B.V.
Peer-review under responsibility of euroCRIS.

Like journal articles, project reports, conference or working papers, PhD theses are results of scientific research work and relevant for current research information systems (CRIS). They contain valuable information on research projects, institutions and experts useful for the management of expertise and skills of persons and organizations. After a short overview we present the support of Common European Research Information Format (CERIF) and review descriptive metadata formats of ETD from existing infrastructures in order to identify elements that can be linked to CRIS. The discussion puts the focus on metadata, on the interoperability between institutional repositories and research systems and on the processing of complementary material, especially of datasets.

2. ETD and CRIS infrastructures

Theses and dissertations are “the most useful kinds of invisible scholarship and the most invisible kinds of useful scholarship” (Suber 2012). Along with journal articles, they are the most important content of open archives. More than half of all institutional repositories listed in the OpenDOAR directory contain theses and dissertations. The academic search engine BASE provides more than 2.7 million ETD via the OAI-PMH protocol. The DART-Europe portal gives access to nearly 500,000 open access research theses from 555 universities in 28 European countries.

In several countries, ETD are processed via specific national or regional infrastructures, frequently connected to local or other repositories. Unfortunately, most ETD systems are developed unrelated to CRIS infrastructures. Linking from ETD to CRIS or accessing ETD directly from CRIS is difficult or impossible. Now, “a CRIS could be seen as dysfunctional if the actual research results were difficult to access by the CRIS users” (Rabow 2009). It makes sense to add links to the full text, and the “desire to create efficient research information systems was (…) an important impetus for the creation of dedicated and standardized publication repositories” (ibid.).

However, linking ETD to CRIS is not a major issue for the development of IR or CRIS. For instance, Rabow describes the results of a 2009 survey conducted by the Scandinavian Nordbib programme on the use of CRIS and institutional repositories. Except for the statement that a growing part of repositories contains an increasing percentage of PhD theses (in 2007, about 75% IR contained theses, and 1/3 had more than 50% of the institution’s output), the Nordbib report does not provide any specific information regarding theses and CRIS. Also, American colleges and universities started to develop “researcher-centric” models of institutional repositories, with roots lying “in the open access publishing movement within libraries” (Younglove 2013). These repositories are designed as a platform for researchers to collaborate with one another, store data sets, and publish their research through the library, and they can be linked to or are even part of research systems. Younglove cites two universities that made their repositories become more like CRIS systems, the Harvard repository DASH and Indiana University's OJS. But again, there is no specific information about theses.

Yet, the challenge of this topic has been identified and addressed elsewhere. Sachini et al. (2010) describe the emerging Greek ETD infrastructure including an ETD repository and insist on the CRIS connection with structured information about authors, organisations as well as research projects which have specifically funded PhD theses.

The UK Embed-Project assessed barriers to acceptance of IR by the scientific community (Harrington & Betts-Gray 2009) and identified the requirements for a system which would “facilitate the more effective management of research outputs from submission to external exposure”. The repositioning of IR within an integrated CRIS was tested. Their conclusions were in favour of a “flexible single submission system providing for one time deposit and uploading to multiple external dissemination outlets including IRs, subject repositories, personal and departmental web pages” and “more training and advocacy to ensure that researchers retain final pre-publication versions of their work where appropriate”. Including ETD in the IR would be “a powerful example of proactive dissemination” leading to “dramatic results” in terms of usage (access statistics) and visibility (contact requests).

The University of Novi Sad (Serbia) developed since 2008 a CRIS called CRIS-UNS based on the CERIF format and compliant with OAI-PMH of institutional repositories and the international ETD-MS format of the NDLTD network (Ivanovic 2012). Their data model contains all elements about theses and dissertations prescribed by the CERIF model (Ivanovic et al. 2012a). The Novi Sad team mapped their data model against the library MARC 21 format (Ivanovic et al. 2011) and published their ontology of theses and dissertations necessary for the creation of a web service that makes the CRIS-UNS metadata publicly available (Ivanovic et al. 2012c). It is so far the best model for the topic of our paper. Partly based on the Novi Sad model, Peponakis (2013) published another ontology of ETD metadata from library records. In particular, his data model shows the “division of a University to Schools and

Departments along with the assumed relationships between persons and theses as well as dissertations (and) the persons' mobility between universities"; allowing for statements like "'Person A' is a man who earned a Master and a PhD at 'Faculty E' and 'Faculty B' respectively, which both are subdivisions of 'University X'. His master's thesis supervisor was 'Person B' and his PhD supervisor was 'Person C', both being female. We also get information about 'Faculty B' which is a subdivision of 'School A' and was established in 1963" (ibid.).

3. CERIF and ETD

The basic features of the CERIF format allows to integrate PhD theses in research evaluation, in particular by means of the core entities *Person* and *Organization*, the results entity *ResultPublication* and different 2nd level entities as *cfCountry*, *cfLang*, *cfFundProg* and so on (Jeffery et al. 2002, Ivanovic 2012). The CERIF semantic layer can distinguish between different types of theses and dissertations and describe a person's role as author or advisor of a PhD thesis. The Novi Sad mapping of CERIF, Dublin Core and ETD-MS shows the way (Ivanovic et al. 2012a,b). Their work clearly indicates that the richest and most exhaustive formats are the ETD-MS developed for the ETD network NDLTD and the library MARC format while the Dublin Core and CERIF are less specific, in particular with regards to other persons and roles than the author of the thesis, alternative title, thesis type, field and discipline and dates. They list 13 metadata about theses and dissertations with their storing method in the CERIF data model, as a link, entity or attribute (author, title, subtitle, keywords, abstract, note, ISBN, total pages, publisher, publication date, URI, thesis type, institution).

They also describe their own CRIS-UNS data model based on CERIF but more exhaustive insofar it integrates 18 additional elements from the library MARC 21 records, for instance advisor, chair, committee member, access rights, name of degree, scientific discipline or date of defense (see Ivanovic et al. 2012a, tables I and V). The Novi Sad project demonstrates the flexibility and adaptability of CERIF for the specific needs of ETD. Yet, all MARC 21 or ETD-MS metadata may not be necessary for research evaluation (such as extended abstract, physical description or even content format). A CRIS including ETD for research evaluation finally may be able to make do with a CERIF less exhaustive than the CRIS-UNS data model.

4. Case studies

4.1. Slovakia

The Central Registry of Theses and Dissertations (CRTD) has been created following the initiative of Ministry of Education, Science, Research and Sport of the Slovak Republic. The primary goal of the registry is to support for science, research and education by manage collection and archiving of final examination works and second doctorate works from universities and colleges in Slovakia (Turňa et al. 2012). All documents from higher education institutions (involved 33 institutions) in Slovakia operating according to the body of laws of the Slovak Republic must be sent to this central registry and subject to an originality check before their defense. This collecting system operates on two levels. The university or college will provide its own local repository and on the second level will forward metadata to the global repository. The CRTD, together with selected internet resources serve as a comparative corpus for the system for plagiarism detection, called also the Antiplagiarism System (APS). Once the work is checked for plagiarism, an output protocol is produced for each work, which serves the exam commission as a tool for deciding about the work originality (Noge 2011). The system has been running since 2009. Its database comprises graduation publications (bachelor, diploma and doctoral) and qualification postdoctoral publications of academic members of all Slovak universities from 2010. The CRTD contains almost 300,000 documents. Expected increase counts around 75,000 documents per year (Noge & Dušková 2013). All documents sent to the register since September 2011 are also published on the crzp.sk portal. Presentation of works to the public depends on the parameters of the license agreement and may include metadata or full text. The CRTD metadata structure follows principles of UNIMARC format. SOAP Web service offering the metadata in MARC XML. The xsd schema contains a description of several entities used in CERIF.

Organisation means university and its basic part (faculty, institute) guaranteeing relevant study programme. They are registered in three levels: University – Faculty – Department. For instance, TNFPTKMTaE means:

TN – Alexander Dubček University in Trenčín,
FPT- Faculty of Industrial Technologies in Púchov and
KMTaE – Department of Material Technology and Environment

Person in ETD has defined following fields: Name, Surname, Affiliation Pers_OrgUnit. Person could also play multiple roles in relation to ETD work: adviser, supervisor, reviewer, opponent, consultant, author. These roles are compatible with CERIF semantics categories. If the document is prepared by more than one person, the percentage of person's contribution in a document can be indicated.

4.2. Iran

In Islamic Republic of Iran, as Iran 1404, the Vision Document or 20-Year Document requires the country should be the first in the region. This has resulted in a national research movement and hence, the number of graduate students and the number of published scientific documents have considerably grown in the recent years (Mehrtash & Fatemi 2012). To make the research output accessible for the scientific community, it is important to index it through a set of standard metadata in which research objects could be retrievable for the other researchers. We note that this is true for both our national publications and also for world publications; however the focus of this section is on our own national dissertations and theses in order to bring maximum visibility and accessibility of them for the researchers. To achieve this purpose, a national CRIS called SEMAT was designed to provide a collaborating environment across institutions by integrating national scientific data. (Khoshroo & Fatemi 2010). SEMAT is only integrating metadata of research entities. It is designed to make the research information discoverable for researchers. However, the full text of the research objects (such as dissertations) is required for scientific work. There is an ongoing effort in Iran, to make individual full text repositories for every research object. In this paper, we show the methodology of making a national repository of full texts of dissertations and theses called "GANJ", the Persian word for treasure. Since GANJ is planned to act as a complete national repository, we started the process for establishing GANJ through SEMAT and CSTIS ("Commission of Science and Technology Information System"). SEMAT, the Iranian national CRIS, is managing research information across all institutions in Iran. On the other hand, CSTIS, a policy making and standard establishing body, is developing standards for metadata management of all research objects in SEMAT and also is responsible to assign national level projects to Iranian institutes (Mehrtash & Fatemi 2012).

The new national ETD repository which stores every ETD along with its full text addresses all the issues mentioned above. All data are registered and managed during the work flow of every individual object in a central registration system. The idea is to register and collect related data of the ETD during the typical workflow of initiating a thesis/dissertation, approval of the proposal and defending the work as illustrated in Fig. 1.

All institutions (including universities and research institutes) are required by CSTIS to setup registration services for every thesis. The GANJ registration system communicates with their online submitting system, once the thesis receives acceptance.

The minimum requirement for GANJ to accept registration of the thesis is: "successful defense". In these points, which are shown by flags in Fig. 1, metadata and full text are being registered in GANJ and thus the scientific information becomes retrievable for researchers. This infrastructure satisfies major quality criteria:

- **Interoperability:** Every IR system with its own workflow sends the data to both SEMAT and GANJ. The initiation of each transfer is by the originating system (i.e. IR).
- **Completeness:** Since every individual thesis is going through the institute workflow, we can be sure about the completeness of the repository. Every thesis is passing by the flags shown in Fig. 1 and therefore it will be registered in the system.
- **Validity:** Every individual thesis is going through a defense session by the defense committee assigned by the institute. This process and the final judgment by the committee are the best validation proofs.
- **Freshness:** As soon as the student's work is being successfully defended, it will be registered and therefore will be available in the repository. This process assures the freshness of the repository and hence, all the current information are being collected in the system. However, it is planned to start the registering ETD from earlier stages such as proposal approval.

- **Homogeneity:** To register the thesis in the system, a web service is designed and implemented. The web service enforces one single format of the data being entered in the system.

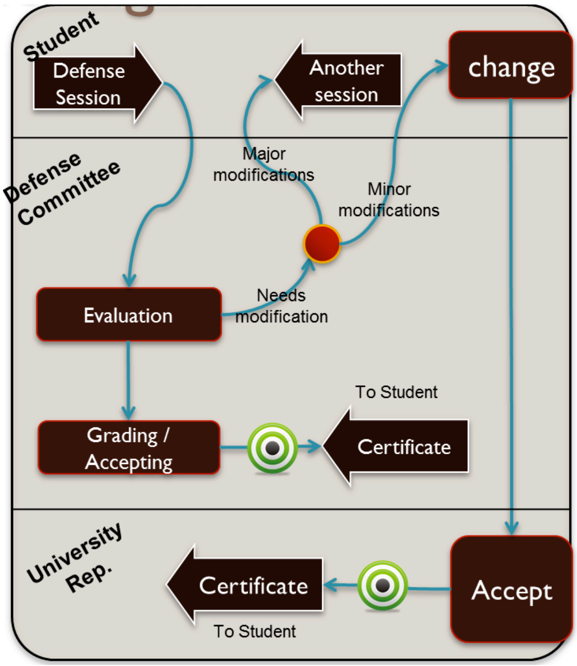

Fig. 1. The typical process of item submission.0

As described before, SEMAT (Iranian national CRIS) is integrating metadata of all research entities. On the other hand, GANJ is responsible to integrate all data related to theses and dissertations including full-texts and metadata. In addition to these two national systems, there are also the institutional repositories or research information systems of every individual institution. The interoperability among these systems is explained below.

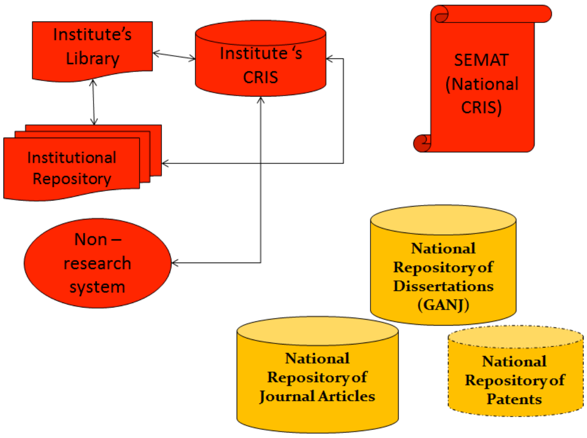
Institute's Library
Institute 's CRIS
SEMAT (National CRIS)
Institutional Repository
Non – research system
National Repository of Dissertations (GANJ)
National Repository of Journal Articles
National Repository of Patents

Fig. 2. The variety of systems using research information.

There are lots of systems including research systems and non-research systems which involve in research information as shown in Fig. 2. For example, human resource system as a non-research system of an institution is holding the information of researchers.

This “enterprise bus system” design acts as the national research hub where every system is physically connected to this bus through Internet as shown by solid lines in Fig. 3.

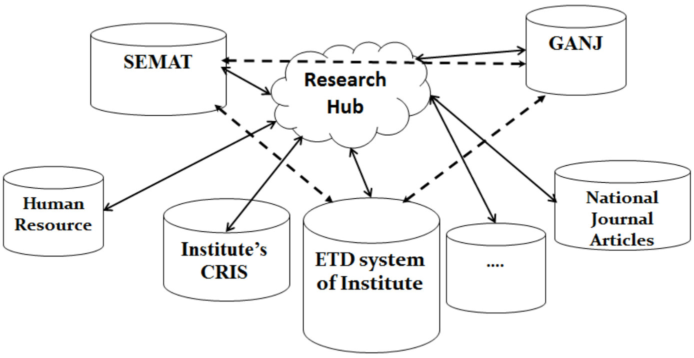

Fig. 3. National research hub.

By having the physical connection between every system and the hub, the logical connections between any two systems is feasible as shown by dotted lines in Fig. 3. Therefore any message generated in the ETD system of an institute is delivered to both GANJ and SEMAT through this hub. The metadata of the article is stored in SEMAT database while the full information including the full-text is stored in GANJ. Once the physical and logical connections established, all systems need to communicate. Such a distributed environment introduces many issues like latency, synchronization and partial failure of one system. A loose coupling using message passing mechanisms has been preferred, with messaging protocols based on xml. Each message is issued by an authorized issuing system which is the origin of the data. Each system sends the message to the hub where there is a message buffer. Whenever the connection between the target system and the hub is alive, the messages are sent to the target system.

4.3. France

For more than 40 years now, the academic union catalogue SUDOC reports all French PhD theses in a MARC format, with records produced by the local university libraries in the central system hosted by ABES and developed by OCLC-PICA (Paillassard et al. 2007). Theses from 1985 on are available via the recent portal Theses.fr which allows retrieval of published theses, theses in preparation, persons (authors, advisors), institutions and topics (Giloux 2012). More and more institutions are switching to the new ETD infrastructure with two connected systems, STEP for theses in preparation and STAR for published theses (Giloux & Mauger-Perez 2008). Both systems use the same XML metadata format called TEF 2.0, a recommendation produced by a group of experts of the French standards organization AFNOR (Ducloy et al. 2006). The TEF 2.0 format contains 75 descriptive, administrative and legal elements compliant with the rich union catalogue MARC format and with qualified Dublin Core metadata, and it integrates items from MODS (Library of Congress) and ETD-MS (NDLTD). Most of the values are controlled.

Actually, approximately 2/3 of the PhD theses are submitted in digital format. All these files are transferred to and preserved on a central server hosted by the French computer centre CINES. The author and/or the local university can deposit the ETD on an institutional or another repository, for open access dissemination but there is no mandatory policy for ETD or open access to ETD so far. More than 42,000 ETD are openly and freely available on the national ETD platform TEL hosted by the French Centre for Direct Scientific Communication CCSD. The format of TEL, as well as of many institutional repositories hosted by the CCSD, has been developed by the CCSD (OAI-HAL) and is compliant with the Dublin Core and the OAI-PMH protocol. Yet, there is no direct link to the rich MARC format or to TEF. For instance, the TEL format requires (with poor control) the elements institution, language, advisor and French key words but not graduate school, field or jury members which are optional elements. Another problem with TEL is that the central repository is not integrated into local institutional workflows. This means that authors sometimes deposit another than the validated version of their PhD thesis. Both ABES and CCSD start to develop features compliant with a CRIS, such as unique identifiers for persons and organizations but there is no national CRIS, and especially in the Higher Education sector very few organizations implemented their own CRIS, while waiting for a recent CRIS initiative led by the Ministry of Higher Education and Research.

5. Discussion

5.1. Metadata

Metadata on PhD theses are produced in different formats and with different degrees of richness and complexity. Basic elements are: author, title, institution, date, advisor, key words. This may be enough for retrieval in repositories but it is not sufficient for CRIS. At least three aspects must be taken into account:

Standards: ETD metadata should be produced and disseminated in a standard format and not in a format specific to an institution or network, in order to facilitate the integration and exploitation by a CRIS. The minimum standard is the Dublin Core.

Richness: PhD theses contain more information about research than minimal metadata often represent. In order to enhance research evaluation by CRIS, ETD metadata could or should include the following elements: author with affiliation, validating institution (university) with full details (town, region, country, type), hosting institution e.g. corporate society or laboratory (if eligible), title, field (discipline), key words, identifier, advisor(s) with full affiliation, members of jury with full affiliation, date (year of defense), research project with start and end dates (if eligible), funding agency (if eligible), protection (confidentiality), dissemination (address). Most of this information is part of usual library or graduate school records but not of IR metadata. Therefore, whenever possible ETD metadata for CRIS should be imported from a local database, an academic union catalogue or a national registry and not from an open repository which often contains poor information about PhD theses.

Quality: Evaluation needs controlled, validated and quality data. Metadata can be very heterogeneous, with different information in the same field and different variants of the same information. Therefore, in addition to the standard and rich format, metadata should whenever possible contain controlled values, based on authority files and curated data, in particular for persons, organizations and fields.

5.2. Interoperability

In order to exploit ETD for research evaluation, ETD infrastructures, repositories and CRIS must connect and exchange data. Needed are compliant metadata and formats, not data silos with specific features. Interoperability problems are partly caused by non acceptance of a standard by the community (Peponakis 2013). All solutions should therefore prefer accepted and already implemented standards, whenever possible. The well documented Novi Sad project can serve as a model, with the implementation of CERIF and the mapping of different formats and data to the CERIF specifications. CERIF appears to be the best choice for flexibility and interoperability. Yet, the format must be extended to integrate rich and complex metadata on ETD, in particular on the semantic layer enabling classifications of entities (types of theses, roles of persons…) and relations between entities (person/organization…).

5.3. Complementary material

Print theses and dissertations have regularly been submitted together with complementary material, such as maps, tables, speech samples, photos or videos, in various formats and on different supports. In the digital environment of open repositories and added value services, these open data could become a rich source of research results and datasets, for reuse and other exploitation. With regards to CRIS, this material can and should be taken into account as research results, along with products or publications – not on the same level as big data from important scientific equipments but as small (dark, personal) data, related to individual work or specific research projects.

This is not a major issue for CRIS but should be considered for further developments, with implications for CRIS formats, workflows, repositories and metadata.

6. Recommendations

Based on the studies and experiences cited in our paper, we will issue five recommendations:

1. Whenever possible, ETD should be part of CRIS.
2. Only validated theses, after defense and controlled by national or local structures, should be taken into account.
3. Metadata should be as standard and as rich as possible and compliant with CERIF and OAI-PMH, even if not all ETD metadata are necessary for research evaluation.
4. Whenever possible, the metadata should be linked to the freely available full text deposited in an institutional or other open repository.
5. Special attention should be paid to complementary material submitted along with ETD, because of their interest for scientists (reuse) and evaluation (research results).

References

J. Ducloy, et al. (2006). `Metadata towards an e-research cyberinfrastructure: the case of French PhD theses'. In DCMI '06: Proceedings of the 2006 international conference on Dublin Core and Metadata Applications, pp. 133-148. Dublin Core Metadata Initiative.

M. Giloux (2012). `Theses.fr - Access to French PhD'. In ETD 2012. 15th International Symposium on Electronic Theses and Dissertations. Lima, September 12-14, 2012.

M. Giloux & I. Mauger-Perez (2008). `A new French circuit for the electronic theses'. In ETD 2008 11th International Symposium on Electronic Theses and Dissertations, 4 - 7 June 2008, The Robert Gordon University, Aberdeen, UK.

J. Harrington & M. Betts-Gray (2009). `The Embed Project: Final Report'. JISC and Cranfield University.

D. Ivanovic, et al. (2011). `CERIF compatible data model based on MARC 21 format'. The Electronic Library 29(1):52-70.

D. Ivanovic (2012). `Software systems for increasing availability of scientific-research outputs'. Novi Sad J. Math 42(1):37-48.

L. Ivanovic, et al. (2012a). `A data model of theses and dissertations compatible with CERIF, Dublin Core and EDT-MS'. Online Information Review 36(4):548-567.

L. Ivanovic, et al. (2012b). `Integration of a research management system and an OAI-PMH Compatible ETDs repository at the university of Novi sad republic of Serbia'. Library Resources and Technical Services 56(2):104-112.

L. Ivanović, et al. (2012c). `CRISUNS ontology for theses and dissertations'. In ICIST 2012 - 2nd International Conference on Information Society Technology, February 29-March 03, 2012, Kopaonik, Serbia, pp. 164-169.

K. G. Jeffery, et al. (2002). `Comparative study of metadata for scientific information: The place of CERIF in CRISs and Scientific Repositories'. In Proceedings CRIS2002 6th International Conference on Current Research Information Systems

P. Juznic (2010). "Grey Literature produced and published by Universities: A Case for ETDs". In D. Farace & J. Schöpfel (eds.), Grey Literature in Library and Information Studies, pp. 39-51. De Gruyter Saur.

M. J. Khoshroo & O. Fatemi (2010). `SEMAT, National Current Research Information System for IRAN'. In CRIS 2010: Connecting Science with Society - The Role of Research Information in a Knowledge-Based Society. 10th International Conference on Current Research Information Systems, June 2-5, 2010, Aalborg, Denmark.

P. Mehrtash & O. Fatemi (2012). `CSTIS, Policy Making Body for National Research Information System in IRAN'. In e-Infrastructures for Research and Innovation. Linking Information Systems to Improve Scientific Knowledge Production. Proceedings of 11th International Conference on Current Research Information Systems CRIS2012.

J. Noge (2011). `Central register of theses and disserations in Slovakia and document originality verification as a centrally provided service'. ProInflow.

J. Noge & M. Duskova (2013). `Central Registry of Theses and Dissertation and the Anti-Plagiarism System as a comprehensive solution at national level'. In Seminar on Providing Access to Grey Literature 2013: The 6th year of the seminar focused on storage and providing access to the grey literature, 23th October 2013 [online]. Prag, National Library of Technology.

P. Paillassard, et al. (2007). `Dissemination and preservation of French print and electronic theses'. The Grey Journal 3(2):77-93.

M. Peponakis (2013). `Libraries' Metadata as Data in the Era of the Semantic Web: Modeling a Repository of Master Theses and PhD Dissertations for the Web of Data'. Journal of Library Metadata 13(4):330-348.

I. Rabow (2009). `Research Information Systems in the Nordic Countries - Infrastructure, Concepts, and Organization'. Report, Lund University Libraries Head Office.

E. Sachini, et al. (2010). `A service-oriented national e-theses information system and repository'. In The 5th International Conference on Open Repositories, Madrid, Spain, 6-9 July, 2010.

P. Suber (2012). Open access. MIT Press, Cambridge, MA.

J. Turňa, et al. (2012). `The system SK CRIS, scientific publications and theses – mirror of Slovak science'. In Proceedings of the 11th International Conference on Current Research Information Systems (June 6-9, 2012, Prague, Czech Republic).

A. Younglove (2013). `Rethinking the Digital Media Library for RIT's The Wallace Center'. D-Lib Magazine 19(7/8).

All web pages accessed in March and April 2014.