The National Repository of Theses: A Short Polish Case Study

Jarosław Protasiewicz[1], Małgorzata Stefańczuk[2], Andrzej Sadłowski[3]
National Information Processing Institute, Laboratory of Intelligent Information Systems
address: al. Niepodległości 188b, 00-608 Warsaw, Poland,
e-mail: [1]jaroslaw.protasiewicz@opi.org.pl, [2]mstefanczuk@opi.org.pl, [3]asadlowski@opi.org.pl

Keywords
electronic theses and dissertations repository, information system, big data

1. ABSTRACT

The aim this study is to outline the main assumptions and challenges that occurred while introducing the national repository of theses in Poland. More specifically, we discuss the legal basis and other conditions of the repository, its architecture, implemented business processes, and some the most interesting technical details and performance statistics. Particularly, we show that a simple datastore based only on a filesystem could be more efficient and cheaper than a NoSQL database, which is quite fashionable nowadays. Presented comments and remarks regarding development and deployment of the information system and selected statistics of the working software may help other parties to improve their repositories or take a decision regarding new solutions.

2. INTRODUCTION

In recent years, many universities and research units decided to preserve the works of their students and researchers in an electronic form. To properly manage data, it requires a dedicated storage with appropriate software. In the literature, such information system is usually named as a repository of Electronic Theses and Dissertations (ETD), which may be considered as the particular case of an Institutional Repository (IR). The difference between them is that the IR covers all possible documents of an institution; whereas, the ETD repository includes only theses and dissertations of students or researchers. We have to note that there are also other types of repositories like disciplinary, aggregation, governmental repositories, etc., but in numbers, most of them are institutional (Yiotis, 2008).

The main purpose of introducing ETDs is to publish students' works (Yiotis, 2008), so they can become easy available for a wider audience. In addition, they may be retained cheaper and longer than in a printed form. We have to be aware that investments in EDTs do not produce any direct revenues but increase access to students' works and allow their wider dissemination (Galea, 2014). An embarrassing problem appearing in universities is plagiarism (Kravjar & Dušková, 2013). The ETDs may be conveniently integrated with anti-plagiarism programs to locate such practices. Nevertheless, this issue is outside the scope of this study.

The works may be openly available on the Internet or only within a local institutional network. There is no clear evidence that worldwide published works have a bigger impact that the others because there is no relationship between their downloads and citations. However, one may observe a local influence, because a thesis becomes discussed again in a local academic environment (Bennett & Flanagan, 2016). Some advocate that an internet repository is not synonymous to open access and the authors should have opportunity to take decision about openness their work (Schöpfel & Prost, 2013). Since some researchers prefer to publish subsequent publications based on their thesis, there are some concerns whether the works openly stored in an ETD are welcome for future publication in journals or university periodicals. The detailed research has shown that most redactors would likely consider such works for publication or with some preliminary check (Ramirez, Dalton, McMillan, Read, & Seamans, 2013).

Many institutions developed their own unique software to manage electronic theses and dissertations but the most worth noting are open source solutions, which are deployed in several places. For

example, about half institutional repositories over the world use DSpace[1] (developed at the University of Southampton) or ePrint[2] (developed at the Massachusetts Institute of Technology) platforms (Ramirez, Dalton, McMillan, Read, & Seamans, 2013). Other important repository platforms are as follows: Virginia Tech University's ETDdb[3], Greenstone[4], Digital Commons[5], OPUS[6], dLibra[7], etc. (Galea, 2014; Richard, 2004; Vijayakumar, Murthy, & Khan, 2006; Yiotis, 2008). Each repository requires an efficient method for data transfer from sources in various locations. Most repositories implement the Open Archives Initiative Protocol for Metadata Harvesting[8] (OAI-PMH) (Ramirez, Dalton, McMillan, Read, & Seamans, 2013; Vijayakumar, Murthy, & Khan, 2006; Yiotis, 2008). Nevertheless, other protocols ale also in use like Really Simple Syndication, Simple Web-service Offering Repository Deposit, the Atom Publishing Protocol, and others (Ramirez, Dalton, McMillan, Read, & Seamans, 2013).

The objective of this work is to briefly outline the main assumptions and challenges, which occurred when introducing the national repository of theses in Poland. More specifically, we discuss the legal basis and other conditions of the repository, its architecture, implemented business processes, and some the most interesting technical details and performance statistics. We have to note that an information system implementing the repository is a part of the information system for science and higher education in Poland (Protasiewicz, Michajlowicz, & Szypke, 2016).

The novelty of this study lies in describing a Polish case study, sharing comments and remarks regarding development and deployment of the information system, and showing interesting statistics of the working software. They may be particularly valuable if we consider possible disasters that could occur during development or maintenance period of ETDs like data loss (metadata, full text, administrative data), backup failures, insufficient security policies, wrong architectures, etc. (Perrin, Winkler, & Yang, 2015).

This paper is structured as follows. Section 2 contains the introduction and related works. Section 3 explains the legal bases and other assumptions of the national repository of theses in Poland. The architecture of an information system implementing the repository covers Section 4; whereas, its business processes are discussed in Section 5. Technical and performance details are included in Section 6. Finally, conclusions and references are presented.

3. THE LEGAL BASIS AND ASSUMPTIONS

In 2014, The Law on Higher Education in Poland (The Act of 27 July 2005) was modified introducing in the Article 167b the national repository of theses. According to this, the repository is led by the Ministry responsible for higher education. It has to include all written master, bachelor, and engineer theses finished after 30th September 2009. A thesis record transferred to the repository must include:

- a title;
- the names of a thesis author or authors, a supervisor, and reviewers;
- the names of a University and its department (unit), where the thesis was defended;
- the exact date, when the thesis was defended;
- a name of a study field of the thesis author;
- a full content of the thesis.

Each thesis must be stored in the repository immediately after a student had passed his final diploma exam. It is the responsibility of a relevant university.

Based on the above, we can conclude that there should exist a central information system implementing the national repository, where all theses would be stored and processed. Each University must have access to the repository. We assume that most Universities should have a local

[1] http://www.dspace.org

[2] http://www.eprints.org

[3] https://theses.lib.vt.edu/ETD-db/index.shtml

[4] http://www.greenstone.org

[5] http://digitalcommons.bepress.com

[6] http://www.kobv.de/entwicklung/software/opus-4

[7] http://dlibra.psnc.pl

[8] https://www.openarchives.org/pmh

repository of theses produced by their students. Thus, the most convenient for the Universities would be permanent integration their systems with the central repository through interfaces like web services. However, some of them may have insufficient resources or only a few students so that integration could be economically unfounded. In that case, the national repository should provide a web application allowing manual operations on data (Figure 1).

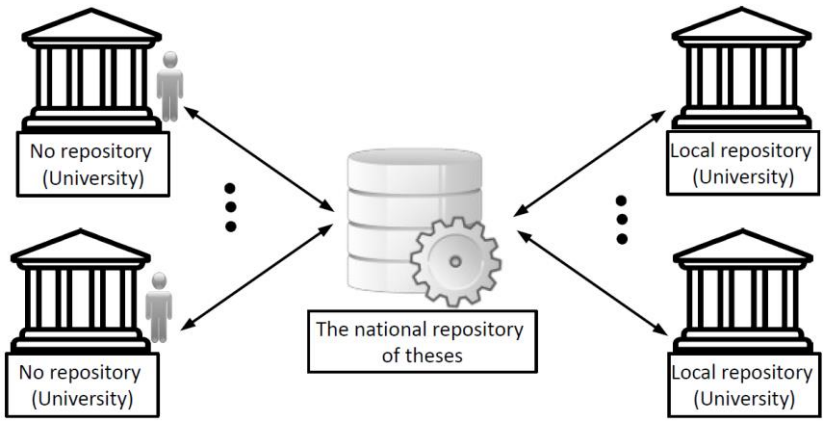

Figure 1. The overall idea of the national repository of theses.
(there are used incons from http://icons8.com)

4. ARCHITECURE AND BUSINESS PROCESSES

The repository architecture is based on the same assumptions like many modern information systems. First of all, it consists several layers separating different levels of data abstraction and various operational purposes, namely (i) a storage tier, (ii) a data access level, (iii) a processing layer, and (iv) a interfaces layer (Figure 2). Likewise in a typical information system, the data access layer provides access to data and, at the same time, it separates business operation carried out in the processing layer from raw data located in the storage layer. In the same manner, the processing layer is responsible for all computations, data validations, and handling clients' requests, and concurrently, it separates customers from accessing data available in the data access layer directly through user's interfaces.

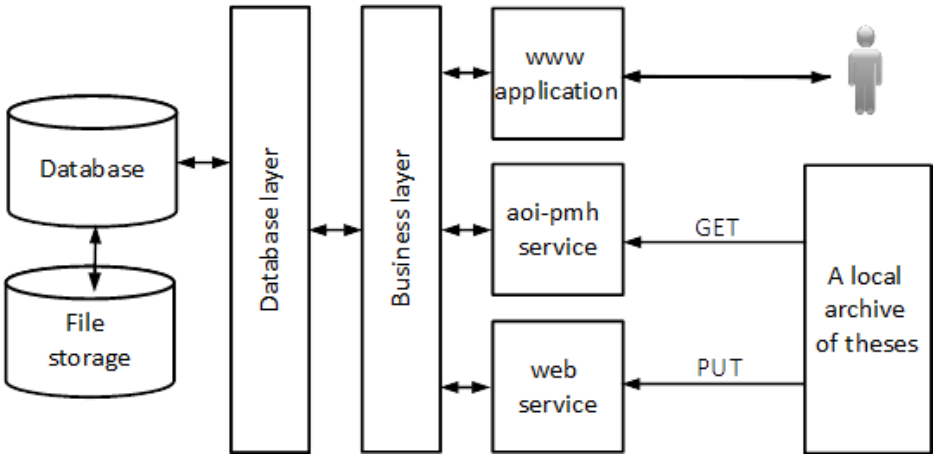

Figure 2. The architecture of the national repository of theses.

On the other hand, we introduce some important improvements of the typical architecture of an electronic theses and dissertations repository. Firstly, the storage layer is composed of a relational database and a file storage, which cooperate with each other. The database persists metadata describing theses, whereas the file storage handles theses' contents. Such solution provides quick search by using database mechanisms and fast access to thesis contents by using native file system mechanisms (for details, see Table 2 in Subsection 5.2). Since the amount of data in such repositories is usually very large, this dual storage solution allows organizing cheap backup processes instead of expensive commercial approaches.

Secondly, the interface layer comprises three submodules, namely (i) a www application, (ii) put and (iii) get services, which handle different types of client's requests (Figure 2). The web application allows manual storing, updating, and browsing data. It is very helpful for small universities with a small number of students or when accidental updates are needed. Contrary to the www application, which is dedicated to humans, both service types are designed to integrate the national repository with the repositories of theses located in universities. A web service based on Representational State Transfer (REST) architecture implements a PUT operation, which means that a client in a local repository actively puts data into the central repository. The client sends a ZIP file comprising an XML file with metadata describing theses and the files containing theses' contents. Opposed to the REST service, a service based on Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) implements a GET operation. It actively harvests theses from the local repositories if they provide an appropriate interface and are registered in the central repository.

5. TECHNICAL DETAILS AND PERFORMANCE

The repository had been developed in 2013[9] and deployed in the fall of 2014[10] as the part of the Information System for Science and Higher Education in Poland (Protasiewicz, Michajlowicz, & Szypke, 2016). Since that time we gained much experience regarding maintenance a vast repository of theses, which we are going to share in this section.

5.1. Implementation technologies

The system has been developed in the Java Enterprise Edition[11] (JEE) technology with the use of Maven[12] tools and the Spring[13] framework. Especially, we used Spring components like Core, Security, Model View Controler, and Data Java Persistence API. An interface layer is based on Java Server Pages (JSP), jQuery[14], and Spring components, which are included in the business and database layers, as well. Additionally, the business layer utilizes a Google library for Java named Guava. Data are stored in a MongoDB[15] database version 2.6.x and a Linux file system.

5.2. The case of MongoDB vs filesystem

Initially, all repository data were stored in a MongoDB database version 2.4.x. However, the performance tests indicated that this database was inefficient due to excessive lags during writing data on a physical storage system. Thus, we tried to upgrade the MongoDB database from version 2.4.x to 3.0.x containing an improved storage engine named Wired Tiger. Unfortunately, we encountered too many obstacles during this operation. In that time, the database software was unstable, and it contained many open issues. Additionally, it was hard to find professionals in this technology, or they were too expensive for us likewise an official support.

The troubles mentioned above led us to attempts in finding other solutions. After preliminary experiments, we decided to use a really straightforward and convenient solution. We left all metadata in the MongoDB database, whereas the entire content of theses was moved to a filesystem. For example, almost 70,000 theses retained in the filesystem were processed about 5

---

[9] The first version of the repository was developed in 2013 by Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw; however, this software was never deployed publicly.

[10] The repository was redesigned and deployed publicly in 2014 by National Information Processing Institute, Poland. More or less, it is the currently running version of the software.

[11] http://www.oracle.com/technetwork/java/javaee

[12] https://maven.apache.org/

[13] https://spring.io

[14] https://jquery.com

[15] https://www.mongodb.com

hours faster than those stored in the MongoDB. Table 1 covers the small excerpt of the empirical results confirming this observation.

Table 1. The comparison of processing times with regard to different data stores, i.e., (i) whole data are stored in the MongDB database version 2.6.x, (ii) the MongoDB covers metadata, and the theses are stored on the Linux filesystem.

Datastorage	Number of theses	Time start	Time stop	Duration
Metadata→ MongoDb 2.6.x Content → Filesystem	67,379	04:11 9:00	05:11 4:44	19h 44m
Metadata and Content -→ MongoDb 2.6.x	69,451	09:30 9:00	10:01 11:05	25h 5m

The fields describing theses are covered in Table 2. As we mentioned before, they are persisted as metadata in the database. Thus, we can easily query the repository by various values of these fields and, based on fast indexes, quickly show the relevant results on the users' interface. The thesis content is loaded from the file system only if it is requested by users.

Table 2 . The fields in metadata, their quantity, and types.

Field in metadata	**Quantity**	**Data type**
Thesis title	1	String
Keywords	0…n	String for each keyword
Thesis authors – first, second, and family names	2…n	String for each name
Thesis supervisors: first, second, and family names	2…n	String for each name
University and department names	1	String for each name
Date of work defense	1	Date
Obtained professional title	1	String from a dictionary, e.g., engnieer, master of science, master of art, etc.
Name field of study	1	String

Additionally, we considered a MongoDB shared cluster, but it required too much hardware in comparison our means. Eventually, we decided to use the upgraded version (2.6.x) of the MongoDB database for metadata and the simple Linux file system for all theses files. In our opinion, this solution is cheap and fast enough for our purposes, e.g., cooperation with a plagiarism system. Nevertheless, the plagiarism issues are outside the scope of this work. We expect influx about 1.2TB data each year and believe that this storage would handle these data without any problems. We have to note that there are no particular reasons for using a NoSQL database like MongoDB for storing metadata. It could be any (relational or object-oriented) database. We still operate the MongoDB database due to too high costs of refactoring to another database.

5.3. Usage statistics

The all statistics presented in this subsection represent the state of the national repository of theses on February 20, 2017. By this date, the repository has covered over 1 million theses and approximately 3.5TB of data. On average, a thesis size is equal to 3.7MB. However, it varies from 2MB to 8MB depending on a university type or its category (Table 3 and Table 4).

Table 3. An average size of theses in respect to university types.

Type	Average thesis size [MB]	Percentage of theses [%]
University	3.021	53
Technical University	8.246	3
Medical University	2.344	5
Other units	3.225	39
Average	3.717	

Table 4. An average size of theses in respect to university categories.

Category	Average thesis size [MB]	Percentage of theses [%]
Public	3,901	79.7
Non – public	3,079	18.7
Catholic	2,112	0.4
Others	2,010	1.2
Average	3,717	

As we mentioned in Section 4, there are three ways of storing data in the repository. Initially, we expected that the OAI-PMH service should be the most popular among clients as it is an open source protocol and widely known over the world. Surprisingly, the most common interfaces are the web application for manual operations by humans and the web service implementing the PUT operation for transferring data from other programs (Table 5).

Table 5. The usage of interfaces to sore data in the repository

Interface	Number of clients [%]	Percentage of theses [%]
Web application	75	42
OAI-PMH service	4	3
Web service (rest)	21	56

Table 6. Types of files stored in the repository

File type	Percentage [%]
Portable Document Format (pdf)	57.0
Micosoft Word (doc, docx)	35.2
Image (jpg, png, tiff)	2.9
Text (txt)	1.1
Rich Text Format (rtf)	1.0
Compressed (zip)	0.7
Open Office (odt)	0.5
AutoCad (dwg)	0.3
Microsoft Excel (xls, xlsx)	0.2
Others	1.5

Usually, theses in the repository are in pdf or doc/docx format. However, we allow other file types. It is important, because a thesis may be composed of many files, e.g., the main dissertation is in pdf/docx, and there are additional images explaining experiments or files covering supplementary projects (Table 6).

6. CONCLUSIONS

In this study, we briefly outlined the main assumptions and challenges, which we encountered when introducing the national repository of theses in Poland. We tried to share our experiences regarding development and deployment of the information system. Particularly interesting is the fact that a data storage based on a filesystem turned out to be better than a NoSQL database. The low popularity of the OAI-PMH protocol among clients is another surprising fact. Presented in this article comments and remarks regarding development and deployment of the information system and selected statistics of the working software may help other parties to improve their repositories or take a decision regarding new solutions. The further work should explore in more details an optimal architecture of an electronic theses and dissertations repository, which must satisfy the criteria of fast information processing, easy datastore expansion, and low maintenance costs.

7. REFERENCES

Bennett, L., & Flanagan, D. (2016). Measuring the impact of digitized theses: a case study from the london school of economics. *Insights*, 29(2).

Galea, R. (2014). Implementation of a repository for electronic theses and dissertations (ETDs) at the University of Malta library (UOML): A case study.

Richard, J. (2004). DSpace vs. ETD-db: Choosing software to manage electronic theses and dissertations. *Ariadne*, (38).

Kravjar, J. & Dušková, M. (2013). Centralised national corpus of electronic theses and dissertations. In *Fourteenth International Conference on Grey Literature*, 95–118.

Perrin, J. M., Winkler, H. M., & Yang L. (2015). Digital Preservation Challenges with an ETD Collection - A Case Study at Texas Tech University. *The Journal of Academic Librarianship*, 41(1), 98–104.

Protasiewicz, J., Michajlowicz, M., & Szypke, M. (2016). A brief overview of the information system for science and higher education in Poland. In *EUNIS 2016: Crossroads where the past meets the future EUNIS 22nd Annual Congress Book of Proceedings*, 233–235.

Ramirez, M. L., Dalton, J. T., McMillan, G., Read, M., & Seamans, N. H. (2013). Do open access electronic theses and dissertations diminish publishing opportunities in the social sciences and humanities? Findings from a 2011 survey of academic publishers. *College & Research Libraries*, 74(4), 368–380.

Schöpfel, J., & Prost, H. (2013). Back to grey: Disclosure and concealment of electronic theses and dissertations. In *The Fifteenth International Conference on Grey Literature:" The Grey Audit: A Field Assessment in Grey Literature*", Bratislava, 2-3 December 2013. Text Release.

Vijayakumar, J. K., Murthy, T. A. V., & Khan, M. T. M. (2006). Experimenting with a model digital library of etds for indian universities using d-space. *Library Philosophy and Practice*, 9(1).

Yiotis K. (2008). Electronic theses and dissertation (ETD) repositories: what are they? Where do they come from? How do they work? *OCLC Systems & Services: International digital library perspectives*, 24(2):101–115.

8. AUTHORS' BIOGRAPHIES

Jarosław Protasiewicz (PhD). – he is an assistant professor at the National Information Processing Institute and the head of the Laboratory of Intelligent Information Systems. He received the Ph.D. degree in computer science at the Systems Research Institute of the Polish Academy of Sciences. His areas of interest include agile project management, software design and development, big data, machine learning, and bio-inspired algorithms.

E-mail: jaroslaw.protasiewicz@opi.org.pl

Małgorzata Stefańczuk (MSc) –she is a business and IT systems analyst with over ten years of professional experience. She is a graduate of the Faculty of Computer Science at the Polish-Japanese Institute of Information Technology and the Faculty of Production Engineering at the Warsaw University of Technology.

E-mail: malgorzata.stefanczuk@opi.org.pl

Andrzej Sadłowski (MSc) – he is a leader of a development team at the Laboratory of Intelligent Systems, National Information Processing Institute. He is a graduate of the Warsaw University of Technology. He is interested in all aspects related to data transfer, big data, and after work, he spends his free time cycling, swimming, and playing volleyball.

E-mail: andrzej.sadlowski@opi.org.pl