Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2014) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

CRIS 2014

Current bibliography research information systems in Poland

Aleksander Nowiński[1]*

[1]*ICM, Warsaw University, ul. Prosta 69, 00-838, Warsaw, Poland*

Abstract

In this article we give a general overview of the current state of the systems for managing current research information in Poland. We focus on information on bibliographic data, which has special situation in the current science landscape of Poland.Management of the publication outcome on the university level in Poland is not satisfactionary, as far not each university have proper database for this purpose. Also usually there is proper procedure ensuring, that publications are properly registered in the system, so systems are incomplete. State level aggregation of the publication information has been done each four years, and only limited amount of information has been collected, so there was no pression to ensure proper quality of the present systems. Currently situation is changing, as there is a new central system developed to aggregate most of the data belonging to typical CRIS system. Part of this system is a Polish Scholarly Bibliography, which will attempt to provide current and complete information about publication outcome of the Polish academic institutions. As the system POL-on is developed to cover most of the parts of the information about research and higher education it covers two more important modules of this field - POL-index - a citation index for regional journals and a repository of the theses to be used for antiplagiarism purposes.

Peer-review under responsibility of euroCRIS.

Keywords: POL-on; CRIS; Polish Scholarly Bibliography, citation index

1. Introduction

In Poland most of the research institutions and higher education is state funded. Actual research is done almost solely in public sector, while private universities and companies usually do not conduct any research on their own. This situation has significant impact on the development of the current research information systems in Poland.

* Corresponding author. Tel.: +48 22 87 49 409; fax: +48 22-8749-115.
E-mail address: a.nowinski@icm.edu.pl

Peer-review under responsibility of euroCRIS.

Almost all universities and research institutes are state funded and they are managed by Ministry of Science and Higher Education. The ministry is supposed to supervise the research, but prior to 2011 it has very limited ability to neither collect significant amount of the data on the topic nor analyze it. Therefore ministry activity was focused on high-level reports prepared by the scientific institutions on demand. Data was collected on the university level and then sent to various offices of the ministry, which had separate databases on their own. Result was inaccessible, often duplicated information, usually inaccurate and outdated.

On the other hand universities and research institutes themselves are usually very conservative, and there is no common agreement upon need of CRIS systems in daily operation and management. Probably lack of funds is important factor which prevents introducing more complex systems to the daily routine. Only most important areas of the university activity, like HR management or student grading are well covered by IT systems, and typically these solutions even in scope of a single university are provided by different companies or developed in-house. An example of such system is USOS[1], developed initially by University of Warsaw system for student management. On the other side university administration is usually more concerned with bureaucratic part of the management and tends to choose solutions like SAP (chosen by two largest universities in Poland, Jagiellonian University and Warsaw University), which are clearly not meeting requirements of the CRIS systems.

2. Central system - POL-on

In 2011 a project to develop country wide central information system for research higher education, POL-on, has been started. This system is designed to store all the data required for the government supervision over science and higher education. POL-on system stores data about the students, scientists, research project, research outcome, assets, special laboratory equipment etc. It has been developed in response to requirements which arose due to reform of the higher education and scientific research funding, as enforcing new rules of employment and material support for the students was impossible without detailed and current information.

As the system becomes partially operational it has great impact on the universities, as amount of data deposited in POL-on was a two orders of magnitude broader than in previous years. This has significant impact on the requirements for the existing and new university systems.

2.1. Model of the data aggregation

POL-on has two basic models of data aggregation. First is based on the imports, where users upload XML files into the web interface, and then files are batch-processed over night. This is typically used in case, when amount of the data is significant, like data about the students, which may contain thousands of records, and universities actually already have systems which collect such data.

Second option is to edit data manually using web interface. Most of the data is delivered to the system using web interface. It especially affects areas like equipment or labs, where amount of data is small, and universities do not maintain systems to collect such data, or decided that cost of adopting existing tools to export data is larger than cost of manual work.

2.2. Metadata standards

During design stage it has been decided that POL-on will aggregate very detailed data, especially about students and the researchers. Unfortunately system was developed in separation from the international CRIS community, and during the requirements building phase the compatibility with international systems was considered unimportant. There was an initial concept of using CERIF XML format for metadata import, but during the project setup only early specification of CERIF 1.3[2] was available, which was found appropriate, and would require a significant amount of extension, far beyond compatibility level. Therefore a decision has been made to use custom XML standard with its own vocabulary and schema and enforce all data providers to adopt it. This is the case for most of the system, which seems rather unfortunate.

3. Polish Scholarly Bibliography

A part of the POL-on system is a module for management of the current bibliographic results, known as Polish Scholarly Bibliography (PBN)[3], which aggregates on single platform outcome of the Polish scientist and scientific institutions. This is one of the key modules of the system, as in Poland research units are evaluated basing mostly on bibliometric basis. This system differs from other parts of the POL-on, as it has both broader audience, and allows data editing not only by dedicated users but by the authors of the publications themselves.

3.1. Impact of the scientific unit evaluation

In Poland each state funded research institution is evaluated every three (or four) years. The evaluation is based on the set of indicators, among whose key role plays evaluation of the publication outcome of the institution. The problem is that traditionally data about publications have been gathered every three years, solely for purpose of this evaluation as a part of the special survey. The bibliography collected with this survey has been analyzed, evaluated, and afterwards discarded completely. This discouraged institutions to care about the quality of the provided data. Second problem with this method of the data collecting is that as the frequency is very low, and each time required metadata is different, institutions tends not to prepare data in the past years. Instead they aggregate data in terms of single action rather than the constant process. This is also done often by unprepared office staff instead of information centers of the universities, and therefore outcome quality is very low. A study has been performed on this topic, which shows that in many cases despite of the existence of the proper system collecting publication outcome data for the evaluation was prepared using other sources[4].

3.2. State of the bibliography systems in the research institutions

As required quality of the reported data a result for the evaluation procedure was low, the result was lack of the professional bibliographic outcome management systems in universities.

Prior to creation of the POL-on system study has been conducted, to analyze state of the bibliographic systems. This study, which was part of the project design phase, was conducted in the end of 2011, and it covered 86 universities. This study shown, that only 60% (50) of the universities has a proper system to register information about researchers' publications. In the remaining cases publications were published as a document on the web page, there were systems for only minor part of the university (like single faculty) or – most typical - there were no information about institution publications available at all, which was typical case for small universities. Less than 25% universities and institutes have satisfactory complete information inside. It was deeply surprising, that sometimes even best universities, like Jagiellonian University in Cracow (typically ranked as best or second best Polish university) may have no central system for this purpose. In case of the existing systems there is clear domination of the one of the domestic vendors - System "Expertus-Splendor"[5], which was used in approximately 40% of the existing installations. , or in-house developed (46%). Few institutions chose professional library software (Aleph ExLibris - 7%).

3.3. Polish Scholarly Bibliography

Polish Scholarly Bibliography (Polska Bibliografia Naukowa, PBN) has been developed as platform to integrate all the current bibliographic data, and obsolete old form of surveys for evaluating publication outcome of scientific institutions[3]. The data aggregation should become a constant process on regular basis.

In the face of fact, that Polish universities and institutes lack complex information systems capable of delivery valid, proven data PBN has to provide not only machine interface, but also a standard Web GUI for users to type information by hand. The benefit is that PBN may be used by smaller scientific institutions as a system to manage bibliographic outcome, without need of investment. It is important especially for small institutes, with outcome of hundred publications a year, which do not have resources to buy and system for they own.

In case of the publication management also another significant problem arises: single publication has typically more than one author and often it is a result of collaboration between institutions. Therefore a single entry in the

database may appear in evaluation data for a number of institutes and persons. It requires special rules apply to achieve consensus on the final shape of the bibliographic record.

As the EU regulations described in recommendation of 17.7.2012 on access to and preservation of scientific information[6] oblige member states to create a policy and strategy for Open Access, Poland is also preparing some steps in this direction. PBN is element of such strategy, as it acts not only as bibliographical database, but as full-text repository as well. Users may deposit texts (of course whenever license is appropriate), and basically use it as a state level open access repository. On the other hand it must be clearly stated, that some licenses for "green path" open licenses restrict full text deposition in the central repositories (and PBN without doubt is such repository), so it has only limited application without update of the state legislation.

All metadata inside the system are open access to the users and are accessible not only through the web interface, but also by specialized HTTP REST API. It is very important part of the solution, as it encourages transparency of the process of the evaluation in the future. It also benefits in re-use of the data collected so far. First users already harvest the repository to use it in their own systems.

3.4. Current status of the system

System officially has finished testing stage and has been adopted by a number of users. There is prepared novelization of the law about research funding, which introduces obligatory deposition of the data into PBN system, to enforce data completeness. This law will be effective since September 2014, as the new academic year will start.

4. POL-index

Another interesting POL-on module is is a new initiative to build citation index for Polish journals - POL-index citation database[7]. This is local citation index, focused on humanities and social science area, which are not well covered in large international databases like JCR or SCOPUS[8]. There is an increasing need to evaluate journals in these areas, and there is no good measure of the journal quality without measuring its citations. POL-index database integrates existing bibliographic databases and then with cooperation of the journal publishers it aggregates complete and correct articles metadata. To achieve satisfactory quality of the results, enough data must be collected. It is very important, especially due to fact, that in humanities there is different citation pattern than in science: citations are sparse (most of the cited works are out of scope of the citation index, like sources or very old publications) and it takes much more time to get item cited. To achieve success deposition of the metadata (including bibliographical references) is an obligatory part of evaluation of the journal - if journal will not deposit required information, it will not be referenced on the official list of the scored journals, and publications in such journal are not valued during official evaluation process.

Within the index machine learning-based algorithms for citation resolution[9] are applied to the collected data, using Apache Hadoop system for processing. This step is done using open source CoAnSys framework[10]. Results of this processing are quite well (approx. 75% accuracy), yet not satisfactory. Therefore last step of this processing is a crowdsourcing step, where users are able to correct the results (corrections have to be approved by system operator). The primary outcome of the database is calculation of the local citation coefficient for the participating journals, but we expect to calculate other bibliometric measures as soon as database will aggregate enough data.

5. Central Repository for Master Thesis

In the begining of 2014 PBN has been extended by adding a new module: Central Archive of the Master Thesis (Centralne Repozytorium Prac Dyplomowych - CRPD) for all universities in Poland. As there is growing problem of plagiarism and dishonest practices of the students buying their thesis in internet some counter action had to be taken. There already exists some companies in Poland offering anti-plagiarism system, and some initiatives are taken by universities themselves, yet in such case most important is to provide broad base of the original works to be used as reference for plagiarism detection algorithms. Up till now it was hardly possible, as thesis in Poland is considered intellectual property of the student and it is illegal to use it without his written permission. CRPD enforces deposition of the thesis, and it is guaranteed that it will be solely used for the plagiarism detection, which is

permitted. On the other hand all universities are required to use plagiarism detection systems validating documents against CRPD. This should significantly improve the detection of the dishonest students and overall quality of the delivered works.

The CRPD system has just finished early development phase and is now in beta-testing stage. It will be obligatory to deposit data in it since October 2014.

This development has a significant impact on the university systems. Although there was a requirement to maintain university level archive of the thesis, it was not strictly enforced. Also form of the archive was not well defined, so some of the universities collected students works on some physical media (like CD), which proven to be unreadable after some time. The new regulation will change this and will put some stress on proper archiving thesis. This introduces some market for solutions compatible with CRPD.

6. Conclusions

The POL-on system is fully developed and it now works as a central CRIS system for Poland. The ministry of Science and Higher Education during years 2011-2014 successfully has built a central system which collects information about virtually all the aspects of the research activities and resources. The system is operational, and now it becomes a basis for decision making and evaluation of the current state of the science in Poland. This will hopefully improve quality of the management of the research in Poland. Currently the information access is limited, yet more and more data is exposed to the public, and often it means not only permission to view the data, but the access using API as well. We hope that it will be used to improve transparency of the science management in the future.

This has significant impact on the systems existing in the institutions, as requirements for the data delivery rapidly changed. Prior 2012 universities in Poland could operate with only few IT systems limited in scope. Since POL-on system has been introduced increased amount of reporting forces universities to improve their IT infrastructure. This is very important especially in the area of publication outcome management, as it has been neglected for a long time. Also some benefits of the situation, especially better understanding of the shape of the Polish research and better management seem to be visible in close perspective.

References

1. Mincer-Daszkiewicz, Janina. "Student Admission System for Warsaw University." *EUNIS, 2004*.
2. Jörg, Brigitte, et al. "CERIF 1.3 Full Data Model (FDM): Introduction and Specification." (2012).
3. Nowiński, Aleksander, Wojtek Sylwestrzak, and Wojciech Fenrich. "Polska Bibliografia Naukowa." *Materiały Konferencyjne EBIB* 24 (2013): 1-7.
4. Marcinek, Marzena. "From bibliography to parametric evaluation of research units-library services in transition." (2012).
5. Expertus-Splendor: http://www.splendor.net.pl/
6. European Commission. Commission recommendation of 17.7.2012 on access to and preservation of scientific information. 2012. http://ec.europa.eu/research/science-society/document_library/pdf_06/recommendation-access-and-preservation-scientific-information_en.pdf.
7. Fenrich, Wojciech, et al. "POL-index—Polska Baza Cytowań." Materiały Konferencyjne EBIB 24 (2013): 1-8.
8. Archambault, Éric, et al. "Benchmarking scientific output in the social sciences and humanities: The limits of existing databases." *Scientometrics* 68.3 (2006): 329-342.\
9. Fedoryszak, Mateusz, Dominika Tkaczyk, and Łukasz Bolikowski. "Large scale citation matching using Apache Hadoop." *Research and Advanced Technology for Digital Libraries*. Springer Berlin Heidelberg, 2013. 362-365.
10. Dendek, Piotr Jan, et al. "Content Analysis of Scientific Articles in Apache Hadoop Ecosystem." *Intelligent Tools for Building a Scientific Information Platform: From Research to Implementation*. Springer International Publishing, 2014. 157-172.