OMEGA-PSIR – A SOLUTION FOR IMPLEMENTING UNIVERSITY RESEARCH KNOWLEDGE BASE[1]

Henryk Rybiński[1], Jakub Koperwas[2], Łukasz Skonieczny[3]

[1]H.Rybinski@ii.pw.edu.pl, [2]J.Koperwas@ii.pw.edu.pl, [3]L.Skonieczny@ii.pw.edu.pl
Institute of Computer Science, Warsaw University of Technology
ul. Nowowiejska 15/19, 00-665 Warsaw, Poland

Keywords
knowledge base, digital library, scientific resources, repository, research management, open science, open access

1. ABSTRACT

In 2010 a nation-wide strategic program SYNAT, aiming at building a scientific information infrastructure in Poland, has been launched. Originally, the program has been scheduled for the period of three years, in due of the program implementation it was extended till the mid of 2014. It was financed by the National Centre for Research and Development (NCBiR) in Poland. A network of 16 academic and scientific partners has committed to implement the SYNAT's objectives in the form of a universal open knowledge infrastructure for the information society in Poland.

The scale of the SYNAT program was unprecedented for the Higher Education Sector in Poland. Beyond the system development, a comprehensive portfolio of research problems has been addressed by the partners (Bembenik et al. 2013; Bembenik et al. 2014). In the view of the limited implementation time, the primary goal of the program consisted in meeting the challenges of global digital information revolution, especially in the context of scientific information.

One of the outcomes of SYNAT is the software OMEGA-PSIR (in the sequel Ω-Ψ^R), designed and implemented by a team of Warsaw University of Technology. We present the software – a cutting edge solution for building a research knowledge base of academic institutions. We present functionality of the system, as well as, sketch some applied AI technologies aiming at providing features attractive for the system beneficiaries. It is shown that although a classical repository is the main part of the system, the essential value of the solution is in providing analytical tools, especially useful for the "research management", but also for the researchers, students, and the university administration. Lessons learned from deploying the software at Warsaw University of Technology and other Polish universities are also discussed.

2. BACKGROUND

The last decade has shown an increased interest of the universities in the systems concerning research data management and access to publicly funded research data. In 2010, a dedicated project, SYNAT, has been launched in order to address deficiencies of the scientific information infrastructure in Poland. The ultimate goal of the planned infrastructure was to ensure the dissemination of the Polish nation-wide scientific achievements, and to improve integration and communication of the scientific community, while leveraging existing infrastructure assets and distributed resources.

The main SYNAT construction has been based on two levels of distributed knowledge bases, with a central database at the highest level, and the university ones at the lower levels. Schematically the two levels of scientific knowledge bases are shown in Fig. 1:

1. the central level (the main SYNAT platform - INFONA portal);
2. the university level (repositories held by the universities).

[1] **Note: The paper has been submitted with the intention to take part in the Elite Award contest.**

The software for building the university level knowledge base has been designed and implemented. By now, it has been already successfully deployed at Warsaw University of Technology (in the sequel WUT), and with its use the university knowledge base has been implemented. In addition, it is currently subject to use for implementing research knowledge bases at other six academic institutions in Poland.

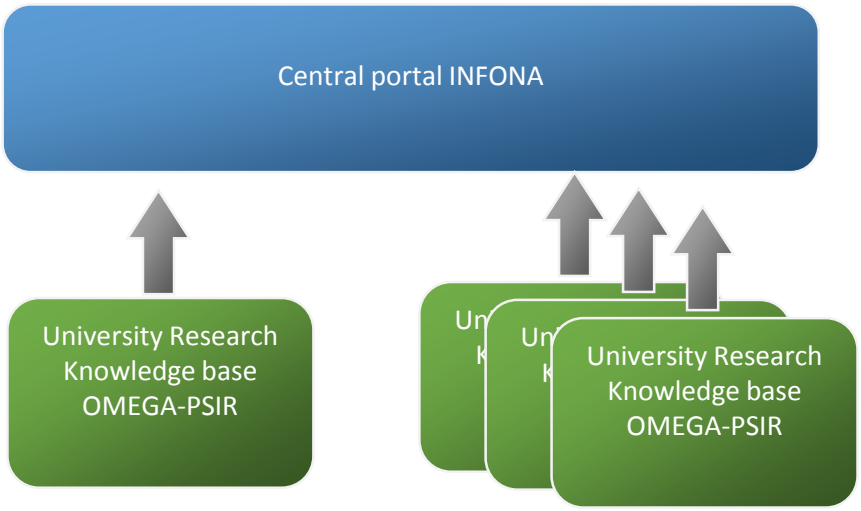

Fig. 1 A general view of the SYNAT network

3. MOTIVATION, GOALS AND KEY ASSUMPTIONS

Observing contemporary information systems dedicated for institutional research knowledge bases, one can see an approach represented by systems like Fedora-Commons or D-space (see e.g. Berman (2008)), which focus mainly on the repository functions, such as storage and indexing of research-related documents, including also aspects of long term durability. It is actually a dominating way for building institutional research knowledge databases. The systems within this approach provide rather simple end-user functionality, mainly limited to browsing and querying the repositories. They are bibliography oriented, usually document-centric ones, and do not provide end users with any analytical functionalities, or with sophisticated presentation capabilities. Additionally, the data acquisition procedures are rather straightforward, based on human work, or harvesting data from well-defined resources.

Although the systems of this kind are in wide use, based on experience of US universities some essential problems have been presented by Davis and Connolly (2007), and Salo (2008). The main criticism of the document-centric approach focuses on a very weak interest of the researchers communities to use such repositories, and can be summarized in one sentence that the institutional repository is “like a roach motel - data gets in but never gets out" (Salo 2008). In particular, it is observed that typical software solutions do not encourage scientists to actively contribute to the institutional repository content:

> *“The institutional-repository software platforms, plagued by innovation-hostile architectures and an ideology-driven rather than user-centered understanding of the problem domain, have been slow to align development with needs. Interested faculty, librarians, administrators, and developers must reframe their approaches to institutional repositories if they are to recover from their current neglect”*

Recently, another approach can be observed – it is researcher-centric and community-oriented approach, like Microsoft Academia, Arnetminer, ResearchGate or Academia.com. Unfortunately such global systems do not cover many of the typical research institution needs. One can therefore observe some initiatives towards building institutional research-centered knowledge base systems. A good example is the Stanford VIVO system (Kraft et al 2010). The VIVO project aimed at creating "Semantic Web-based network of institutional ontology-driven databases to enable national discovery, networking, and collaboration *via* information sharing about researchers and their

activities". Still though, many prominent Stanford researchers cannot be found in the system, probably because too much effort (and cost) should be paid to the database maintenance.

Yet another solution has been offered recently. It is a commercial system PURE proposed by Elsevier[2]. It is an institutional solution, and to a large extent it simplifies the maintenance processes. In general, the idea of building the Ω-Ψ^R platform has emerged from similar motivations. However, as the PURE technologies are not public and are quite expensive, we have focused on elaborating ours. A significant requirement imposed on our system was to make it free, open source, fully customizable and localizable, so that it can be fully adjusted to varying local university conditions.

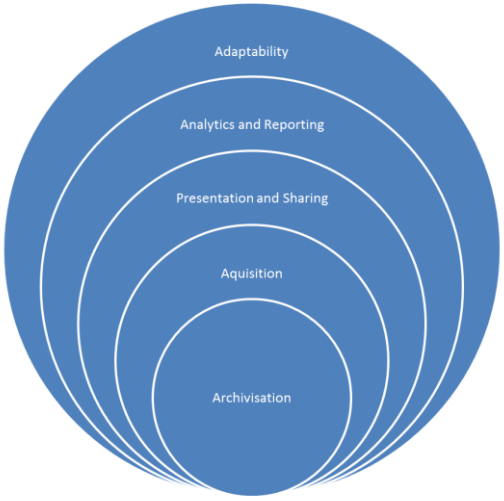

Fig. 2 System requirements

We concluded that the successful software for building institutional research knowledge base has to integrate various, sometimes conflicting, needs of different user groups. It should be beneficial for, and motivating to, quite different user groups, including (but not limited to) researchers, students, university strategic management, administration, librarians. What is very important, it should guarantee low maintenance cost, on the other hand should be as much user friendly as possible. These were the key assumptions for developing the Ω-Ψ^R software.

As the result of the analysis, a multi-level structure of requirements was defined for the system, as illustrated in Fig. 2. The core requirement refers to the needs of archiving the published material, but additional forms of scientific activities of the researchers should be covered by the system. The consecutive important requirement is to provide mechanisms for acquiring data. In this respect, in order to minimize the maintenance costs, the system should be able to acquire data from Internet, as well as, by means of crowd-sourcing. Special requirements refer to the presentation and sharing issues. In particular, special importance has been given to the functions of promoting researchers, university units and informal research teams. Then, the requirements referring to analytical information, strongly related to the presentation requirements, expressed the needs for implementing data mining and knowledge discovery algorithms in order to present the most successful researcher individuals and groups, discover the research maps of the units, and provide statistical information on dynamically changing research potential of individual researchers, informal cooperating groups, or organizational units. The next section presents the functionalities of the system being results of the requirements.

4. MAIN FEATURES AND FUNCTIONALITIES OF OMEGA-PSIR

The main idea for Ω-Ψ^R was:

1. to build tools on top of a classical repository for acquiring data from various sources, and for simplifying the repository maintenance procedures;
2. to provide a number of analytical researcher-centric functionalities, in order to make the system attractive to the users.

Many problems have been solved with the tools of artificial intelligence and text/data mining. In particular, we concentrated on data acquisition from WWW, along with extracting information from the retrieved pages, and then building the knowledge base with the extracted facts. Additionally, in order to improve the quality of the functionalities of the system we concentrated on semantic enrichment of acquired data and facts by automatic indexing and classification of objects, and the presentation of knowledge extracted from the repository data.

[2] http://www.elsevier.com/online-tools/research-intelligence/products-and-services/pure

Repository oriented functions

With any defined object in the knowledge database it is possible to predefine various "digital attributes". The digital attributes are devoted to store digital objects, which then are accessible by a unique "object identifier". The text objects are subject to indexing, so that the index for full text retrieval is built automatically with the new objects added to the database. Also the updates of text documents are automatically reflected in the indexes. The knowledge base can be seen as a network of interlinked objects. For the flexibility reasons, all the object structures are definable by means of XSD definitions, extended by some extra constructs. The definitions contain relationships between the types. From the XSD definition the system automatically builds the forms for data entry, search and result list presentation (Fig. 3).

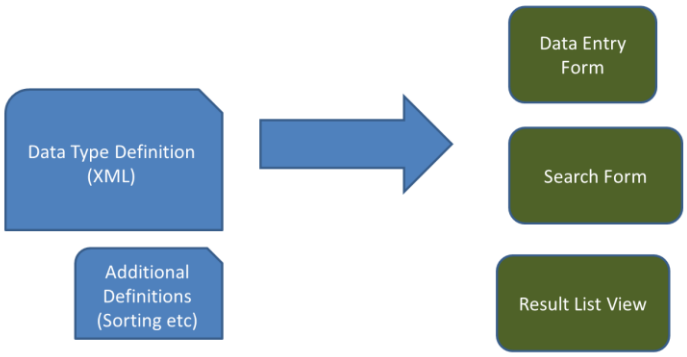

Fig. 3 Defining object types and corresponding interface elements

An important feature is that the repository preserves "historical values" of linked objects in the course of changes (e.g. if an author changes a name his/her publications can be searched equally by the old name and the new one). At WUT the main object types that have been defined are: *researcher*, *publication* (with a number of subtypes), *patent*, *thesis* (with the subtypes BSc, MSc, PhD), *project*, *project_document*, *researcher_activity*. In the near future we expect to add other new types, such as *experiment_data*, *benchmark_data*, *software_tool*, etc.

Acquisition functions

A specialized acquisition module Ψ^R (Platform for Scientific Information Retrieval) has been developed to acquire data from the Internet in a reliable way and import various formats. The main capabilities of the subsystem refer to acquiring data from WWW, along with extracting information from the retrieved pages and building the knowledge base with the extracted facts, and performing semantic enrichment of acquired data and facts by automatic indexing and classification of objects.

Advanced artificial intelligence tools have been implemented for performing a pipeline of acquiring new data, enriching them semantically and integrating within the knowledge base (see Koperwas et al (2014b). To perform the pipeline the following components have been implemented:

- Web Search Module that finds resources related to the scientific world on the Internet. This module is triggered by users' actions or Scheduler that periodically invokes predefined searches on Web;
- Classifier and Extractors modules. The modules are used to decide whether found resources are of a given type, e.g. conference homepage, and extract information for the found resources;
- Disambiguation Module, which assigns publications to the proper researcher record from a set of people with the same first and last name.

It is worth noting that during the implementation of the knowledge base at WUT at the end of 2013, some 20000 bibliographic descriptions of the most important publications of the WUT researchers have been acquired from WEB.

Independently of the AI based means, involvement of researchers directly into the data acquisition process was presumed as a psychologically important factor for achieving data completeness. Bearing in mind a possible drop down of data quality, unavoidable for such approach, a variety of specialized tools guarantying high level quality of the acquisition process have been developed in addition (Koperwas et al 2014a).

Presentation and Sharing Functions

The presentation aspects were subject of our particular interest. Special focus has been put on presentation and promotion of researchers, as well as, university units, and informal research teams. To this end, advanced algorithms have been elaborated and implemented for:

1. tagging the researcher expertise and visualizing it by a cloud;
2. discovering experts in a given domain, based on the research achievements registered in the knowledge base;
3. finding the networks of cooperating researchers.

These algorithms have been used for implementing the main knowledge base functionalities, such as e.g. looking for experts, presenting the researchers profiles, showing the achievements of the university units (institutes, faculties, departments, etc), and generating reports. A sequence of the steps with functions using the algorithms mentioned in (2) and (3) above is illustrated in Fig. 4.

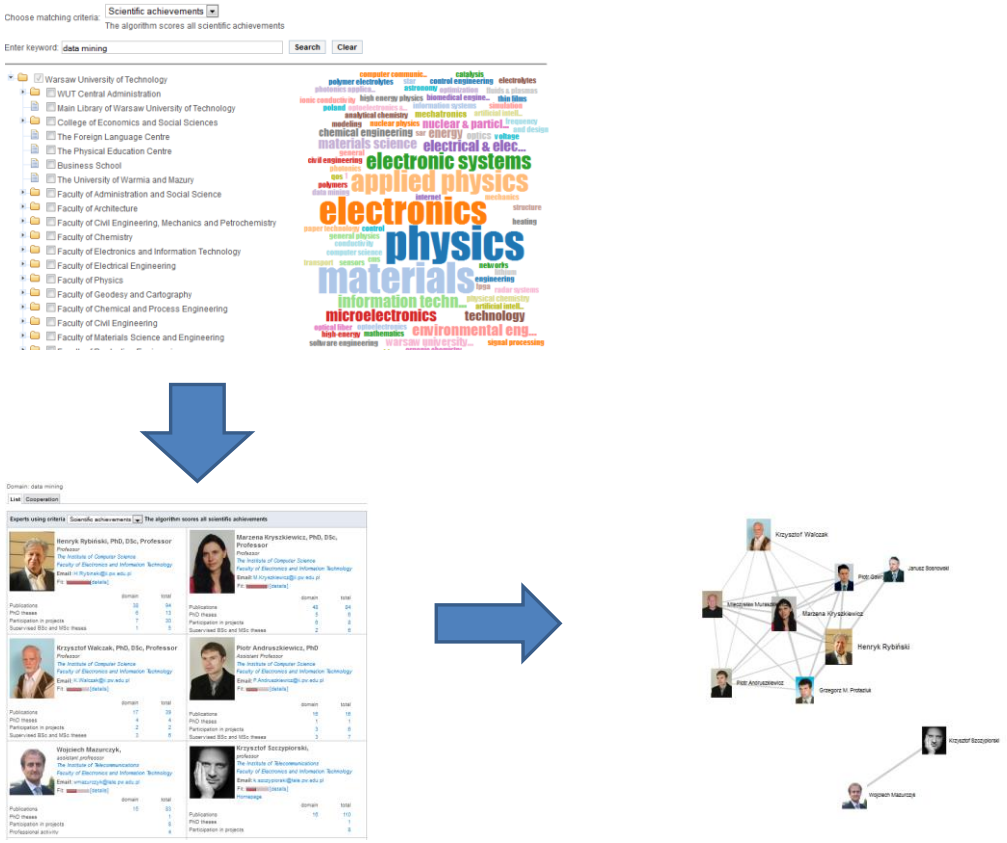

Fig. 4 Cloud of tags and experts search

Another important issue is that the system takes advantage of storing the knowledge base as a graph, so one can easily navigate between various objects. In addition, fairly standard functionalities, such as building a query, presenting search results of bibliographic data, etc., are provided with highly ergonomic and customizable GUI, with semantic support, and various sorting, reporting and exporting methods available, accompanying all the result screens. For the promotional reasons screens provide means for integrating with social networks (Facebook, ResearchGate, etc.).

The exporting functions make possible to provide data in typical formats, including the ones of publishers (e.g. MODS), and those preferred by researchers (BibTeX). The system can communicate with other systems by means of OAI-PMH interface[3], but also SOAP and REST.

[3] A general idea for the repository was to take into account the paradigms of Open Access.

The presentation mechanisms are researcher-oriented with the purpose not only to present but to promote the achievements of the researchers.

5. IMPLEMENTATION OF KNOWLEDGE BASE AT WUT – IMPACT AND BENEFITS

The Ω-Ψ^R system has been built iteratively over the period of three years. It was installed in production environment in its early development stages – at the beginning only at Institute of Computer Science, with functionalities limited to the basic repository functions. In the course of the development of new functionalities the system was providing more functions, and was delivered to wider range of users, first, in 2011 at Faculty of Electronics and Information Technology, and then in 2013 to the whole Warsaw University of Technology[4], still being subject of further development.

Such approach caused that the system was confronted with its users form the very beginning, and the developers were confronted with real user needs, so that when the system was finally ready to be shared with other universities in the form of a complete Ω-Ψ^R package, it was already mature, well-tested and well-documented.

Already in 2013 the functionality of Ω-Ψ^R went beyond the typical functionality of institutional repository, so it had chances to become a central knowledge source for all types of the WUT research activities, the more that, due to applied intelligent tools (acquisition tools, reporting functionalities), the maintenance efforts of Knowledge Base are essentially reduced, compared to the typical solutions. The process of moving the system to the University level was already simpler, as the team had experience with organizational and training issues at the faculty level. In addition, by 2013 we had at our disposal a number of means to mine web for the publications of remaining faculties, and whenever possible to import data from local faculty repositories, web pages, or even files. It is worth noting that the most important and valuable bibliography (some 20000 records) has been harvested from web in 2 weeks. The remaining 10000 historical records have been imported from legacy databases.

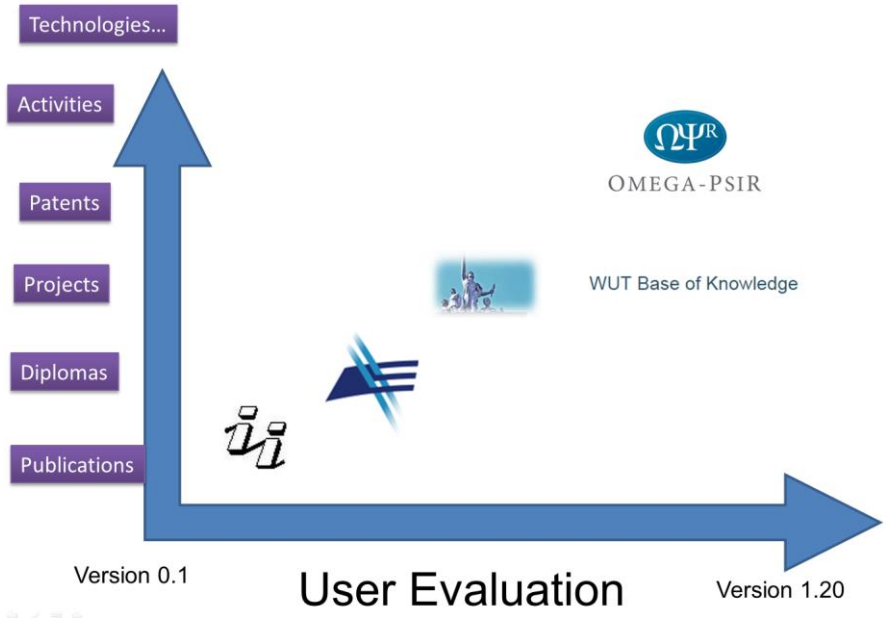

Fig. 5 Evolution of the Ω-ΨR system

As mentioned already, the groups of the system users and beneficiaries are very heterogeneous. As for the internal users, the following groups can be distinguished:

1. Researchers

[4] http://repo.bg.pw.edu.pl/index.php/en/

2. Students (graduates, undergraduates);
3. University administration;
4. Scientific bodies (faculty councils, senate, promotion commissions, etc.);
5. University top management, responsible for research strategies.

As a matter of fact, at the beginning all groups of users were rather skeptical, but the main skepticism was coming from the researchers group. In the course of the knowledge base development the researchers could immediately observe their profiles, so that they gradually turned to be more and more involved in the process of the database maintenance. The main trigger for the staff involvement was the fact that they have noticed correlation between the way the system was presenting their own profiles, and their achievements in the repository. Usually, they were not satisfied with the automatically generated expertize cloud, which was clearly the result of missing publications in the database. Getting familiar with the integration of repository functions, visualization of research, and reporting the university staff became the first beneficiary of the research knowledge base.

Now the system is subject of integrating with other university systems, as well as, with national systems in Poland. Due to the service oriented architecture it is a fairly simple process. In particular, the system is now integrated with Student System USOS, and with the financial systems for research projects. It also integrates resources from all the local repositories and bibliographic databases into one central repository. Due to its analytic functions the system has been easily integrated with staff evaluation and promotion system. It serves already as the main reporting tool for the university authorities. Last but not least, it becomes a source of knowledge about the research teams at the university, and it refers equally to all the users groups, starting from the top management responsible for building research strategy, through project leaders recruiting teams for the research projects, and ending with PhD students looking for their potential supervisors. An immediate result is the process of integrating the researchers groups sometimes from thematically quite far faculties.

Also the efforts to disseminate the University achievements to the external communities start bringing positive results. Also for the external users, the role of the system is multifold. The system integrates various functions, but the main ones are:

1. to provide a complete and up-to-date information about the research areas of the University researchers, and their strength to the potential external partners for building scientific cooperation links;
2. to provide means to the governmental authorities concerning the research potential of the University, and the currents achievements, *inter alia* for the evaluation and assessment reasons;
3. to provide a complete and up-to-date information about the research areas of the University researchers, and their strength to the international evaluating bodies.

The usage statistics of WUT Knowledge Base shows an increasing interest from visitors from all around the world, especially from Western Europe and North America. It is expected that those effects could be enforced with time.

6. THE APPLICABILITY OF THE PROJECT TO OTHER INSTITUTIONS

While building the Ω-Ψ^R software, one of the more important requirements was the flexibility of the system and its adaptability, so that developing new features and changing business rules should be possible by system administrator without involving programmers. In particular, it was planned from the beginning that the system should be:

1. easy to extend (with new data types, views, validation rules, etc.);
2. capable of handling differentiated faculty-specific requirements (for example the procedural ones), even within the same running instance;
3. easy to install and adopt to other universities, also with other interface language

Referring to p. (1) above, a number of scripting tools has been elaborated. As a result, a lot of functions can be expanded or even developed without any need to change the main code. In many cases it can be done in a declarative way, like, e.g., the mentioned above possibilities to define new types of data structure (see Fig. 3) along with the accompanying validation rules, data entry forms and search screens. The system administrator can define custom reports and statistics. Other

advanced scripting options refer to definitions of access privileges and data protection, or even the ways of ranking publications or the researcher expertise.

The system can be embedded in different web pages of the university. For example, at the university level it can be used as a central knowledge base, whereas at the faculty level it can be used for presenting the achievements of the faculty staff. This way the system can serve as a centralized repository for the whole university, on the other hand, each faculty or department may want to promote their achievements on their local homepages and provide a customized look and feel (also by means of colors, and layout). In addition, at the lower level one can provide local reporting styles, local statistics, or export options.

All the flexibility features, as well as the service oriented architecture of the Ω-Ψ^R software make it possible to adopt the system to the needs of other universities. The process of distributing the system in Poland has started. In September 2014 the system has been presented to a group of leading universities in Poland. It turned out to be very attractive for the universities. There are already some 20 requests for providing means for the test evaluation of the system.

The system is available for free of charge, with the access to the software source code. There are already 9 academic and research institutions testing the system, and adapting to their needs. In addition to WUT, yet another university has already implemented its research knowledge base. Interesting enough, they were able to do it without too much support from our side.

It is worth noting that the system is multilingual, so various interface languages can be implemented and the system can be fairly easily adapted to the requirements of universities in other countries.

7. FURTHER DEVELOPMENTS

One of the lessons learned was that with building an information system, for the first glance looking as a fairly typical one, we have encountered many interesting real life research problems in such areas like knowledge acquisition and discovery, text mining, or information retrieval.

While the practical goals of the SYNAT project have been achieved, within this the Ω-Ψ^R platform has been successfully implemented at WUT as the research knowledge base, and now, it is subject of implementing at other universities in Poland, we are going to continue the development of the system. Two types of work are planned: on one hand we would like to take advantage of the existing system flexibility and add more functionality using the system options; on the other hand we can see a lot of research possibilities that can essentially influence the system functionality and quality.

Within the first track (implementation oriented) we can already enrich the existing system with interoperability with other national and global scientific information systems (e.g. ResearchGate, Google Scholar, Web of Science, Scopus, etc.). Special tasks towards this direction are already planned. Other fairly easy task is to add the functionality for looking for “rising stars” (some research in this direction has been already performed. We also have already tools to enrich the journals database with information harvested from the publishers’ sites, e.g. concerning the information about call for special issues.

Within the second track (research oriented) we can see that the already built repository of scientific publications, mostly in English, is quite heterogeneous in terms of the covered research areas, and as such, it provides a lot of challenges. Special emphasis will be put on semantic cross-lingual search, giving rise to a more symmetric retrieval for English and Polish, i.e., giving similar results for queries regardless of the language. Some work in this direction has already been started (Krajewski et al 2014). In addition, our existing web mining tools aiming at discovering knowledge about conferences still require some more research.

8. REFERENCES

Bembenik R., Skonieczny Ł., Rybiński H., Kryszkiewicz M., Niezgódka M. (eds.) (2013). Intelligent Tools for Building a Scientific Information Platform: Advanced Architectures and Solutions, *Studies in Computational Intelligence*, *vol. 467*, 2013, Springer, ISBN 978-3-642-35646-9, 548 p.

Bembenik R., Skonieczny Ł., Rybiński H., Kryszkiewicz M., Niezgódka M. (eds.) (2014). Intelligent Tools for Building a Scientific Information Platform: From Research to Implementation, *Studies in Computational Intelligence*, *vol. 541*, 2014, Springer, ISBN 978-3-319-04713-3, 290 p.

Berman F (2008). Got data?: A Guide to Data Preservation in the Information Age. *Communications of the ACM, Vol. 51 No. 12*, pp. 50-56

Davis P.M., Connolly M.J.L. (2007). Institutional repositories: Evaluating the reasons for non-use of Cornell University's Installation of DSpace. *D-lib Magazine, Vol. 13, No. 3/4*

Krafft DB, Cappadona NA, Caruso JCR, Devare M, Lowe BJ, Collaboration V (2010). Vivo: Enabling National Networking of Scientists. *Proceedings of Web Science Conference,* 2010, pp. 509-518

Koperwas J., Skonieczny Ł., Kozłowski M., Rybiński H., Struk W. (2014a). University Knowledge Base: Two Years of Experience, in: Intelligent Tools for Building a Scientific Information Platform: From Research to Implementation/Bembenik R. [et al.] (eds.), *Studies in Comp. Intell.*, vol. 541, 2014

Koperwas J., Skonieczny Ł., Kozłowski M., Andruszkiewicz P., Rybiński H., Struk W. (2014b). AI Platform for Building University Research Knowledge Base. **In**: Foundations of Intelligent Systems. Proceedings of ISMIS/ Troels A. [et al.] (eds.), LNAI, vol. 8502, 2014, Springer, pp. 405-414

Krajewski R, Rybinski H, Kozlowski M. (2014). A seed based method for dictionary translation. **In**: Foundations of Intelligent Systems. Proc. ISMIS/Troels A. [et al.] (eds.), LNAI, vol. 8502, 2014, Springer, pp. 415-424

Salo D (2008). Innkeeper at the Roach Motel. *Library Trends, Vol. 57,* No. 2, pp. 98-123

9. AUTHORS' BIOGRAPHIES

Prof. Henryk Rybinski

http://repo.bg.pw.edu.pl/index.php/en/r#/info/author/WEITI-45f977de-460e-4ca2-a67f-3da6c240b7f9/?tab=main&lang=en

Prof. Henryk Rybinski leads Institute of Computer Sciences, Warsaw University of Technology. His main research interest is in intelligent information systems, semantic web, data/text mining, natural language processing and knowledge representation. His current research is concentrated on using text mining techniques for knowledge discovery from text data. He has published more than 130 scientific publications in the area of information systems. For some 35 years Prof. Rybinski has been conducting projects for building information systems for many international bodies (i.a. FAO, UNESCO, UNEP, IFRC, IUCN).

Dr Jakub Koperwas

http://repo.bg.pw.edu.pl/index.php/en/r#/info/author/WEITI-bbda4208-5c68-4329-9882-2899d85cfd52/?tab=main&lang=en

Jakub Koperwas, PhD, is an assistant professor at Institute of Computer Sciences, Warsaw University of Technology and lead consultant and partner in IT consulting company - Sages. His research interests are data mining of semi-structured data, especially for bioinformatics and distributed data mining. He has published 10 scientific publications in the area of information systems. He provides software development lectures for students of Warsaw University of Technology.

Dr Łukasz Skonieczny

http://repo.bg.pw.edu.pl/index.php/en/r#/info.seam?id=WEITI-fa2564c9-3b69-4d0e-b03e-64bc6f279911&lang=en

Łukasz Skonieczny, Ph.D, assistant professor at Institute of Computer Sciences, is one of the main developers of the Ω-Ψ^R system. His research interest is in database systems, data-, text- and web-mining, graph theory and web development. He has in his record 10 scientific papers and, 3 edited books. He participated in a bunch of research projects, and cooperated with many institutions, *inter alia* France Telecom, Samsung, UNEP, FAO, IUCN.