Polish Platform of Medical Research with OMEGA-PSIR - an Open Science approach

Jakub Koperwas[1], Łukasz Skonieczny[1], Renata Sławińska[2], and Henryk Rybiński[1]
[1] Warsaw University of Technology, Poland
[2] Wroclaw Medical University, Poland
jakub.koperwas@pw.edu.pl, lukasz.skonieczny@pw.edu.pl, renata.slawinska@umw.edu.pl, henryk.rybinski@pw.edu.pl

Abstract

With the IT technology development, the Open Science movement, aiming at making scientific research and dissemination of its output accessible to all levels of society, amateurs or professionals, evolves from large bibliographic data, addressed mainly to the scientists, through Open Access movement, along with open scientific repositories and archives, recently ending with integrating institutional repositories (IR) with Current Research Information Systems (CRIS). Some newer IT solutions go further, extending systems by functionalities enabling dissemination of multitudinous aspects of research activities. OMEGA-PSIR, developed at the Warsaw University of Technology (WUT), combines functionalities of IR, CRIS and Research Profiling Systems (RPS). Now, adopted by more than 34 scientific institutions in Poland, the system shows usefulness not merely for the university management people, not only for the research community, but also for the society. A good example of implementing Open Science ideas in Poland is the Polish Platform of Medical Research, which is a unique project that covers a number of institutional OMEGA-PSIR implementations at medical universities and research institutes, and provides a central database that integrates the information from the local systems. In the paper we discuss various aspects of Open Science, referring to the successful implementation of the Polish Platform of Medical Research.

1 Introduction

The first systems dedicated to research information management were institutional repositories (IR). Their main focus was to collect and share the open access research outputs such as publications or PhDs. Due to the fact that they were focused on open access texts, the contribution to fairness of research information is undeniable. However as the open access files are just a fraction of all publications and

the publications are just a fraction of all the research activities or outputs, their true impact was not high.

New possibilities to enhance the Open Science perspectives came with the times of increasing popularity of the Current Research Information Systems (CRIS), sometimes also known as Research Information Management System (RIMS). CRIS systems play an important role in the University IT ecosystem.

One of the main CRISes mission was to gather all research information related to University, so as to ease administrative processes, such as employee evaluation, reporting to authorities, project pre-award administration etc. The initial goals for the CRIS deployments (and it may still be observed in some institutions), were intranet systems fully dedicated to reporting purposes.

Straightforward development was to use the collected data to display the university activities to the world via publicly accessible research portals that fetch the data directly from CRIS. With a researcher-centric approach of the Researcher Profiling System (RPS) it has given rise to RPS/CRIS mixed class systems.

A recent stage of this evolution was a reflection that the functionality of institutional repositories (IR) can be easily integrated within CRIS, enabling access to full texts and research datasets from one all-in-one CRIS/RPS/IR platform. Such approach not only increases the power of reporting, but improves information retrieval, and essentially increases visibility and accessibility of research outputs due to obvious synergies, and what is quite important, without any extra exploitation cost.

Such an approach was adapted by the OMEGA-PSIR system, developed at Warsaw University of Technology (WUT). The system has been already deployed by near 40 scientific institutions in Poland. It combines functionalities of IR, CRIS and RPS. The motivations for incorporating within OMEGA-PSIR functionalities of IR, CRIS and RPS have been discussed by Rybinski et al (2015) Rybinski et al (2017). According to the surveys performed by the Main Library of WUT (see Kubrak 2018) the system shows usefulness not merely for the university management people, not only for the research community, but also for the society.

A good example of implementing Open Science ideas with OMEGA-PSIR is the Polish Platform of Medical Research (PPM). It is a unique project -- it has involved several institutional OMEGA-PSIR implementations at medical universities and research institutes. Each institution has deployed a local version of OMEGA-PSIR, additionally a central database has been implemented in order to integrate all the information resources from the local systems.

In the paper we discuss various aspects of Open Science, referring to the successful implementation of the PPM.

2 Polish Platform of Medical Research

2.1 The Main Challenges

An idea of building a common medical platform aimed at disseminating the research results of Polish medical universities has appeared in the libraries of medical universities in Poland in the middle of the second decade of this century. As a result, at the end of 2017 a project aiming at building Polish Medical Platform has started. It was a joint initiative of seven medical universities of Białystok, Gdańsk, Lublin, Szczecin, Warsaw, and Wrocław (the leader of the project), and one research institute, namely Nofer Institute of Occupational Medicine. The IT part of the project has been performed by the consortium of Warsaw University of Technology and SAGES.

The project was co-financed by European Union within the Operational Programme Digital Poland. According to the project document: “The main goal of a project was to build a platform for the purpose

of presenting and promoting scientific achievements and research potential of medical research institutions."

The central platform was expected to showcase and promote the scientific achievements and research potential of the project partners in the areas of medicine, pharmacy, dentistry, public health, occupational safety and health, ergonomics and health protection. More specifically, it should gather in one place, previously scattered information about medical scientific publications, ongoing research, obtained patents, research equipment and profiles of scientists from the institutions implementing the project. Additionally, in the future, any institution interested in cooperation should be able to join the platform.

In addition, one of the conditions was that each partner should have an institutional system with functionalities typical for IR, CRIS, and RPS. The subgoals referring to the traditional paradigms of Open Science could be seen as requirements for each local system, as well as the ones for the central platform. Some of them mainly referred to the IR functionality:

- ensuring open access to the medical document resources (repository of scientific publications, grey literature, theses, teaching material),
- ensuring open access to research data (repository of research data and other literature documents).

Additionally, a number of subgoals that were specified, could be classified as more specific to the CRIS and RPS functionalities:

- aggregating all research-related information in one platform (scientific achievements, research potential, research outputs etc),
- creating a database of experts in various fields of medicine (scientists' profiles),
- increasing the accessibility of resources collected by Project Partners for the visually and hearing impaired by adapting and digitally presenting data in accordance with WCAG 2.0 standards (currently WCAG 2.1),
- improving the openness of shared resources by increasing their availability according to the five Star Open Data scale.

Clearly, this general formulation of the goals can be performed in various ways, and there is no formal way to assess which approach is closer to the idea of open science. Nevertheless, by proposing a specific IT solution one can implement various functionalities that make the final result better suited to various open science paradigms.

While creating the institutional installations, the main problem was to fill the local systems with data that was stored in various institutional databases, usually in different formats. Bearing in mind that ultimately all locally stored data should be integrated in one central database, it was a very difficult task to unify the standards of different local approaches to data categorization schemes.

Yet another problem was the scale of the project, as in a very short period all the installations should have been working. This required an extra effort, but on the other hand it has shown positive sides of the OMEGA-PSIR software.

2.2 Central Platform

The idea of integrating all the information resources of the institutional CRIS systems under one central platform was quite a challenging task. As mentioned above, a very important, though very difficult issue was a variety of the standards for categorization of data. As all the data were planned to be loaded to a central database, it was necessary to set a common categorization scheme. To this end a

decision was taken to prepare a top-level categorization, leaving the local categorizations almost unchanged.

Having implemented 8 local CRIS installations, the work on the central platform has been started. In this case the main challenge was to design and implement a dedicated functionality that enables the OMEGA-PSIR instances to communicate each other and exchange the data in both directions:

- from the institutional sites to the central one to send in order to send new records,
- from the central site to the instituional ones to send necessary changing data about partnering institutions (e.g. new researchers), so that the cooperation information could be up-to-date.

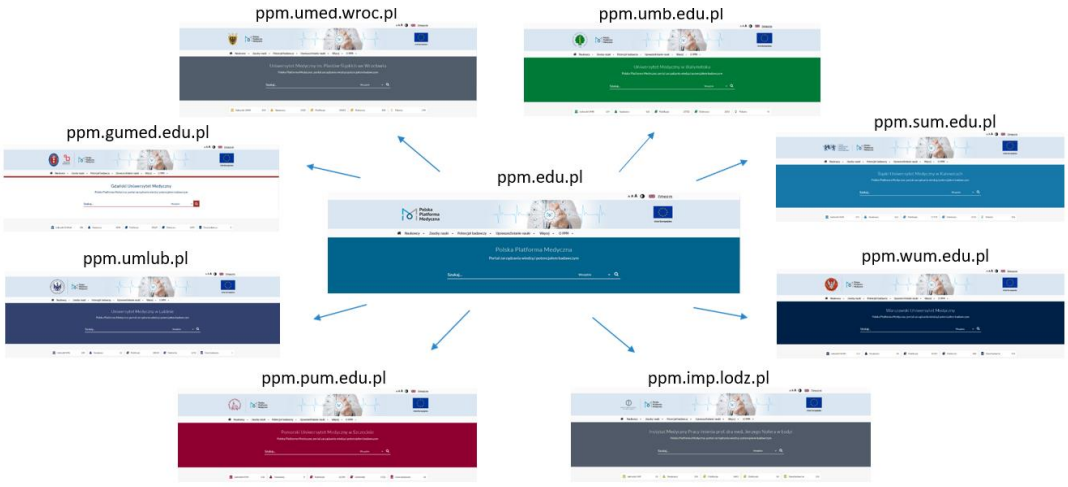


Figure 1 Institutional CRISes building Polish Medical Platform

Figure 2 Institutional CRISes on their way to join Polish Medical Platform

The main feature of this functionality is that every institution can define its own exchange strategy, indicating which data types are subject of exchange with the central platform, and how often, which local records should remain unchanged while importing data from the central system, and more in detail, which data elements within a given object type are allowed to be changed.

The initial resources collected and shared via Polish Medical Platform were:

- 229 243 metadata records consisting of
    - information on publications, PHD theses and research data,

    - research potential (patents, projects, grants and grant applications, research equipment, laboratories),
    - information about events, awards and distinctions, activities, multimedia.
- 20 364 open access documents, stored in the system, consisting of
    - 16,364 journal articles
    - 1 637 chapters from monographs
    - 1,518 doctoral dissertations
    - 376 objects with research data
    - 33 patents

The collected metadata meet the requirements of 5-th star requirement in 5* open data model.

The collected files were manually adjusted to meet the requirements of the WCAG 2.1 / PDF- UA requirements.

2.3 PPM: Living platform

A great challenge was to make the PPM the living platform. The number of data is steadily increasing. The project started with 229 243 metadata records of all kinds, now only the number of publications exceeds 300 000 records. Soon, two new partners will join, and the number of publication records is about to be doubled.

The key to success was to build the local instances before the data has to be gathered at the central system. A real challenge with the institutional CRIS systems was not only to migrate data (and avoid duplicates), but also integrate them with other systems working at the universities, such as, among others:

- local bibliographic and/or library systems from different vendors or built in-house,
- local diploma repositories,
- local ERP systems.

In addition, all the institutional CRISes had to be integrated with the national CRIS system, mainly in order to simplify the reporting of the universities to the national authorities. All this work was worth the effort as currently partners focus on their local needs and the data is seamlessly flowing to the central platform.

Moreover, the integration with various external systems was performed, mainly in order to reduce the database maintenance costs and avoid manual data entry. It referred to the global databases, such as WoS, Scopus, PubMed, SherpaRoMEO, PlumX, Altmetric, ORCID, Google Analytics, European Patent Office, European Databases for Project Calls. A specific integration related to the medical research referred to the MeSH thesaurus.

Currently two more institutions, namely Medical University of Lodz and Jagiellonian University Medical College are on their way to connect to the Polish Medical Platform. Two other research institutes that use OMEGA-PSIR as well, Institute of Mother and Child and Maria Sklodowska-Curie National Research Institute of Oncology who are also adopting OMEGA-PSIR are the next candidates to join the platform.

3 Open Science solutions in PPM

3.1 FAIR principles

The FAIR paradigm as presented by Wilkinson (2016) defines four principles: findability, accessibility, interoperability, and reusability. The rules were defined as follows:

Findable: The first step in (re)using data is to find them. Metadata and data should be easy to find for both humans and computers. Machine-readable metadata are essential for automatic discovery of datasets and services, so this is an essential component of the FAIRification process. […]

Accessible: Once the user finds the required data, they need to know how they can be accessed, possibly including authentication and authorisation. [...]

Interoperable: The data usually needs to be integrated with other data. In addition, the data needs to interoperate with applications or workflows for analysis, storage, and processing. [...]

Reusable: The ultimate goal of FAIR is to optimise the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings.

In the next subsection we discuss how the paradigm is reflected in PPM.

3.2 FAIR principle in research information systems

When we look at the level of the University, the good installation of the CRIS system makes open science a reality. All the research potential, such as staff, expert groups, laboratories, equipment, all research outputs such as patents, products, publications, etc., and activities, such as research projects are visible to the world, and if deposited with full text, they may be downloaded.

However, when we look at the problem from the perspective of the whole science of a given country or a single domain, such as medicine, the situation is more complicated. Can we really say that CRIS systems make the research information FAIR?

Findability and Interoperability: Provided that one wants to find a significant results in given discipline or search for an expert team, she/he must scan all available institutional CRISes and deal with the problem that data range and meaning are not standardized across installations. That definitely does not comply with the findability principle for humans. Machines are in a slightly better situation as sometimes repositories are registered in OpenDOAR or other registries and expose the data with OAI-PMH, however still lack of real data unification and identifiers in DC format makes it impossible to combine the data from different repositories. Those who decided to expose the data in CERIF format are again in a slightly better situation as CERIF itself forces the datasets to be almost 4th star level. Nevertheless, the standardization of the data is not yet fully reached especially in domain specific context.

3.3 FAIR principle in research information systems

An important problem distinguished in FAIR is accessibility data. However, to our knowledge nobody postulates accessibility of repository information to disabled people. All public websites in EU shall follow the WCAG guidelines. It is a relatively easy goal to achieve for a website however the content deposited in a repository is more difficult. Repositories are full of PDF files that are created in various templates in various time periods. Typically, there is no access to the source of those documents so they can be converted easily. So the research information is neither findable nor accessible for physically challenged people.

The conversion to WCAG format of PDF files is not trivial task. From technical point of view it comes down to preparing PDF/UA file - a PDF file conforming a few additional requirements. From

practical, end-user point of view it means that it should be correctly readable by screen readers such as NVDA. This includes the appropriate reading order, alternative texts for figures, document structure for navigation. All of these not provided in a typical PDF file and are hard to be automatically determined. This comes from the fact that even the simple processing of PDF files is cumbersome. It is a format primarily intended for displaying and printing content. Each page of the document is described independently so the text does not flow from page to page. Each page is a sequence of instructions of the type "render an element in a given format in a given place on the screen", meaning that the logical order of these elements does not need to be the same as the order in which they appear in documents. This makes the extraction of coherent text fragments difficult and there are, in principle, no universal methods of this type. In the extreme case, when the document is available only in the form of an image (e.g. a scan), the only possibility of recognizing the structure of a document is the use of image analysis methods, supported by machine learning methods.

During initial implementation of the PPM most of this work has been done manually by a set of annotators. PDF files were manually corrected using the popular PDF editing tool and inserted into PPM along original versions. Additionally, it was made possible for the user to report accessibility problems regarding these files.

That consumed a lot of time and introduced concerns about the future and continuity of this initiative. Therefore another UE fund (POIR.01.01.01-00-0600/20 "A tool that enables automatic adaptation of text and text-graphic documents to accessibility standards") has been raised to develop AI tools for automatic adaptation to accessibility standards. The purpose of this project is not only to bring automated conversion for CRIS systems but also to run a dedicated general purpose SaaS platform for parties that would like to convert their text resources or individuals. At the moment of writing this paper the tool is already working and may be integrated with other systems, the platform is to be implemented within next months.

To achieve the task, we developed ai models for document structure recognition, reading-order recognition, table structure recognition, figure types classification, and alternative text description for tables, images and figures. All of these were initially trained on the data that was gathered during the deployment of PPM, that is medical papers and phd theses which were manually adjusted to WCAG standards. Then we made an effort to extend the set of documents as it turned out to have some deficiencies in many aspects, e.g. in alternative texts. The works were carried out in two ways. First, a team of annotators manually corrected selected documents from the PPM collection and - for diversity - prepared WCAG versions of documents other than scientific papers. Secondly, we automatically generated a WCAG layer for such documents from PubMed repository for which there was a description in the so-called JATS (Journal Article Tag Suite) format. Due to the publishing policy of many medical journals, authors of publications must provide, in addition to the pdf file, a structured, textual form of the document that is logically very similar to the structure required by the WCAG layer. This additional JATS file, however, is not directly related to the pdf file and often there are content inconsistencies between them, but with careful handling it can be used to recreate the structure in the original PDF file.

By combining information from two files, an initial description of the logical structure of the document was automatically determined, and additionally verified by the annotators.

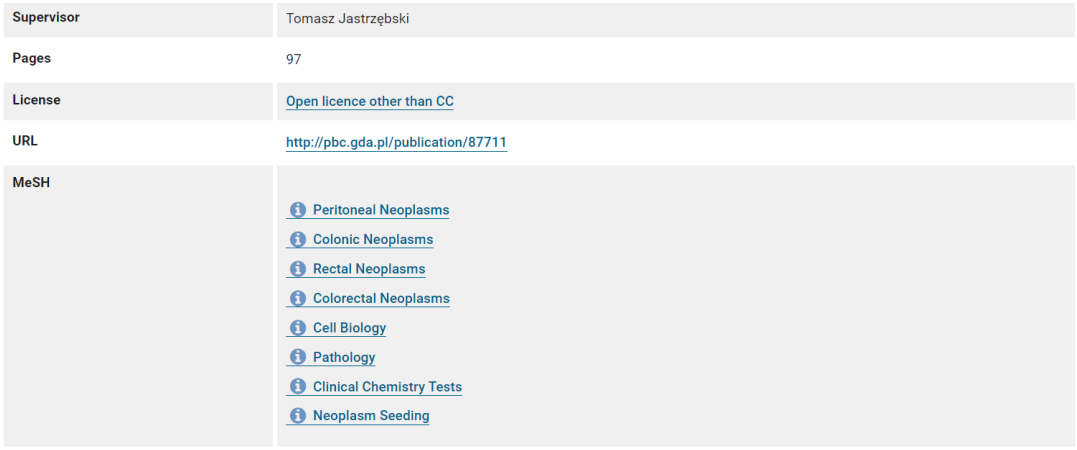


Figure 3 MeSH in PPM

3.4 Five stars open data

A complementary paradigm, Five Stars LOD, should be included to the Open Science requirements. It was introduced by Berners-Lee (2008). It defines 5 levels of openness:

- 1 star is granted when the data is available
- 2nd star is granted if the data is in machine readable format
- 3rd star is granted for non-proprietary format
- 4th star is given for semantic format such as RDF
- and 5th star is given for linkage to external data.

All the metadata is automatically enriched so it meets the requirements for 5th star in 5th star open data scheme. All possible entities such as:

- Authors
- License types
- Journals
- And many others

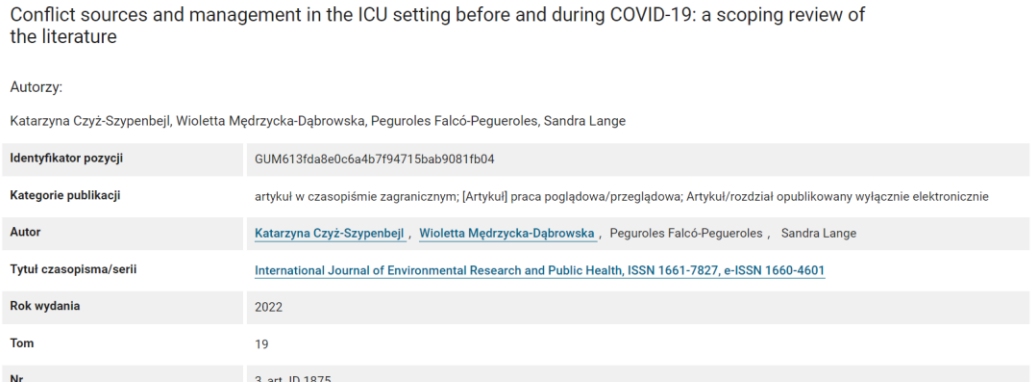
are linked to external schemas such as wikipedia, creative commons, orcid and many others. The one of the most important elements for the 5star description is the linkage to MeSH ontology that is of a special importance to medical sciences.

3.5 Open Science = more information

As OMEGA-PSIR integrates CRIS IR and RPS, it is quite natural that the system can show not only bibliographic data. As shown by Rybinski et al (2017) quite an essential traffic to University research knowledge bases refers to the following:

- units profiles
- research teams
- researcher profiles

Beside the resources the great output of the project was building the researcher profiles out of collected data.



Conflict sources and management in the ICU setting before and during COVID-19: a scoping review of the literature

Autorzy:

Katarzyna Czyż-Szypenbejl, Wioletta Mędrzycka-Dąbrowska, Peguroles Falcó-Pegueroles, Sandra Lange

Identyfikator pozycji	GUM613fda8e0c6a4b7f94715bab9081fb04
Kategorie publikacji	artykuł w czasopiśmie zagranicznym; [Artykuł] praca poglądowa/przeglądowa; Artykuł/rozdział opublikowany wyłącznie elektronicznie
Autor	Katarzyna Czyż-Szypenbejl , Wioletta Mędrzycka-Dąbrowska , Peguroles Falcó-Pegueroles , Sandra Lange
Tytuł czasopisma/serii	International Journal of Environmental Research and Public Health, ISSN 1661-7827, e-ISSN 1660-4601
Rok wydania	2022
Tom	19
Nr	3, art. ID 1875

Figure 4 Five stars in PPM


```
  {
    "@id" : "https://ppm.gumed.edu.pl/info/e/GUM613fda8e0c6a4b7f94715bab9081fb04/",
    "@type" : [ "fabio:JournalArticle", "schema:Article" ],
    "doi" : "https://dx.doi.org/10.3390%2Fijerph19031875",
    "creator" : [ "https://ppm.gumed.edu.pl/info/e/GUMdab7bcd34960414086673 67a070ae824", "https://ppm.gumed.edu.pl/info/e/GUM326e092d81124425afb8b087274850c0",
"https://ppm.gumed.edu.pl/info/e/GUM73cccaf8e3564c1c9307a8afb2d3f71a", "https://ppm.gumed.edu.pl/info/e/GUM23430a621f2648678f26af4c30e18db8" ],
    "language" : "http://id.loc.gov/vocabulary/iso639-2/eng.rdf",
    "title" : "Conflict sources and management in the ICU setting before and during COVID-19: a scoping review of the literature",
    "author" : [ "https://ppm.gumed.edu.pl/info/e/GUMdab7bcd34960414086673 67a070ae824", "https://ppm.gumed.edu.pl/info/e/GUM73cccaf8e3564c1c9307a8afb2d3f71a",
"https://ppm.gumed.edu.pl/info/e/GUM23430a621f2648678f26af4c30e18db8", "https://ppm.gumed.edu.pl/info/e/GUM326e092d81124425afb8b087274850c0" ],
    "isPartOf" : "https://ppm.gumed.edu.pl/info/e/GUM99b23b8fd6194163a92e67a62396c7a1/",
    "license" : "https://creativecommons.org/licenses/by/4.0/"
  }, {
    "@id" : "https://ppm.gumed.edu.pl/info/e/GUM73cccaf8e3564c1c9307a8afb2d3f71a/",
    "@type" : "schema:Person",
    "seeAlso" : [ "https://search.crossref.org/?q=0000-0001-8377-4893", "http://europepmc.org/search?query=%28AUTHORID%3A%220000-0001-8377-4893%22%29", "https://www.scopus.com/results/authorNamesList.uri?
orcidId=0000-0001-8377-4893" ],
    "sameAs" : "https://orcid.org/0000-0001-8377-4893",
    "affiliation" : "https://ppm.gumed.edu.pl/info/e/GUM94788158b0d04055ba51c994049e9af7/",
    "familyName" : "Mędrzycka-Dąbrowska",
    "givenName" : "Wioletta"
```


Figure 5 RDF in PPM

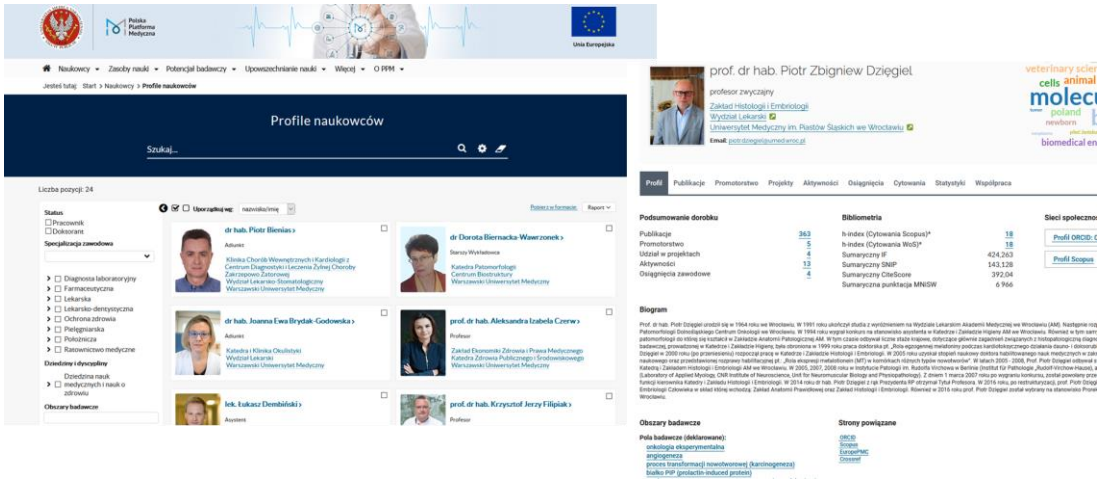


Figure 6 Researcher profile in PPM

4 Summary

Once each university has the CRIS system installed the next step in building the ecosystem for Open Science is the aggregation/federalisation of institutional CRISes into nation-wide or domain-wide systems. The Polish Platform of Medical Research is a good example of a successful approach where the medical institutions agreed on sending data from local CRISes to a central platform that shares unified information about publications, PhD's, projects, research datasets, patents and many others. A great effort was made to unify the vocabularies in local institutions to achieve the goal. All the metadata is 5* LOD sets available to be fetched in RDF or CERIF. All the PDFs are shared in accessible format, manually prepared by project staff. The project was finansed by EU funds granted to a Consortium of 8 Medical Institutions. Currently more medical institutions will join the platform soon. All local installations and central platforms were built on the Omega-PSIR CIRS system by a Consortium of Warsaw University of Technology and Sages company. Further possibilities for developing this platform are being seeked, for instance integration with other databases such as Bio Banks are considered. The project results built the foundations for another EU funded project for AI tool for automatic adaptation to accessibility standards. The tool may be integrated with OMEGA-PSIR as well as other CRIS, Institutional repositories or other document-oriented systems. A separate SaaS platform for PDF conversions for wide range of users will be provided next year. Out of the project another valuable initiative is being prepared due to a funding granted to Jagiellonian University - a tool for translation of MeSH ontology to other languages.

5 References

Berners-Lee T. (2008). Linked open data. See also: http://www. w3. org/2008/Talks/0617-lod-tbl, 281:29, 2008.

Kubrak W (2018). Operation of CRIS system in the constantly changing environment. *EuroCRIS* 2018.

Rybiński, H., Koperwas, J. J., & Skonieczny, Ł. (2015). OMEGA-PSIR–a solution for implementing university research knowledge base. *EUNIS* 2015

Rybiński, H., Skonieczny, Ł., Koperwas, J. J., Struk, W., Stępniak, J., & Kubrak, W. (2017). Integrating IR with CRIS – a novel researcher-centric approach. *Program-Electronic Library and Information Systems*, *51*(3), 298–321. http://doi.org/10.1108/PROG-04-2017-0026

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3(1), 1-9..

6 Author biographies

Jakub Koperwas, PhD

Jakub Koperwas, PhD, is an assistant professor at Institute of Computer Sciences, Warsaw University of Technology and CEO in Sages company. He is Open Science evangelist, co-author of OMEGA-PSIR Current Research Information System. He led many CRIS deployments for Polish Universities including Polish Platform of Medical research. His research interest is in distributed data mining and in information systems.

Łukasz Skonieczny, PhD

Łukasz Skonieczny, PhD, assistant professor at Institute of Computer Sciences, is one of the main developers of the OMEGA-PSIR system. His research interest is in database systems, data-, text- and web-mining, graph theory and web development. He has in his record over 20 scientific papers and 4 edited books. He participated in a bunch of research projects, and cooperated with many institutions, inter alia France Telecom, Samsung, UNEP, FAO, IUCN.

Renata Sławińska, MSc

A graduate of Pedagogy at the University of Wroclaw, postgraduate studies in library science and information science at the University of Wroclaw and in information systems at the Wroclaw University of Technology. An employee of the Main Library of the Medical University in Wroclaw since 1986, in the years 1997-1998 deputy director, since 1999 he has been the director of the library. Member of many university committees, author of publications in the field of scientific information, bibliometry, library teaching, editor of the UMW Library information service, member of the Editorial Board of the Forum of Medical Libraries. Project Leader Polish Platform of Medical Research.

Henryk Rybinski, PhD, DSc

Henryk Rybinski is full professor at Institute of Computer Science. In 2008-2016 he was Director of the Institute. Since 2010 he is responsible for the development of Knowledge Base software for WUT. His current research is focused 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).