Towards interoperability between Omega-PSIR and OpenAIRE

Extended Abstract

Introduction

Omega-PSIR has been developed at Warsaw University of Technology in 2010-2012 as a software platform for building institutional knowledge base covering all activities concerning research run at the university.

The main motivation for building the system has been presented by X1 et al (2015) and X2 et al (2017). It was to provide a functionality of such institutional system that would cover needs of as wide group of users as possible. The problem, however, is that the users groups and beneficiaries of institutional information systems are very heterogeneous. On one hand the needs of researchers, students focus on the access to the gathered resources, on the other hand the needs of the university managers are mainly clustered around the information needed for defining research strategies and developing the university research potential. Yet other requirements may be addressed by the university administration, responsible for providing reports about projects run by the university. Last but not least the whole university community is very much interested in propagating university successes to the scientific world. To this end, and bearing in mind discussions concerning weak acceptance of institutional repositories (IR) by the research community (Davis & Connolly (2007), Salo (2008), or recently Bull & Schultz (2018)), and on the other hand, the importance of IR for the research institutions we have decided to combine the functionalities of IR and CRIS within one system. In addition, in order to increase interest of research staff in building the system we have taken into account functionalities of Research Profiling System (RPS).

As a result, a cutting edge solution for building a research knowledge base of academic institutions has been developed, combining functionalities of the CRIS systems with IR and RPS. By now, the system works at some 10 Polish universities, serving there as a typical institutional repository, but additionally providing the management of the universities with a complete view of research achievements, as well as reporting to the national authorities through the national level CRIS system.

Yet another important function of the system becomes nowadays dissemination of the gathered knowledge *via* recognized international systems. Among them especially important role is given to OpenAIRE[1], which is the European infrastructure for Open Science, and since a few years is moving from a publication infrastructure to a more comprehensive platform that covers all types of scientific output (Principe et al (2014)). OpenAIRE is expected to harvest information from CRIS systems by means of the CERIF-XML format, which has been agreed as the standard format for information retrieval from individual CRIS systems[2].

In the paper we discuss possible scenarios for enabling interoperability between Omega-PSIR and OpenAIRE. We have performed already some experiments towards implementing the support for CERIF-XML in the Omega-PSIR system[3]. We will present modelling tools in the system OMEGA-PSIR and explain the adopted data model. In particular we explain how relationships between object in the model are presented. Special attention will be put on preserving history in the object life. Then, we

[1] https://www.openaire.eu/

[2] The data model of the OpenAIRE infrastructure is CERIF-compliant

[3] In January 2018 Warsaw University of Technology has signed a letter of intent with euroCRIS in which it declared the will to implement the support for CERIF in Omega-PSIR in the near future.

will analyze similarities and differences between Omega and CERIF data models and propose possible mappings. We also present results of initial experiments concerning conversion between the models by means of intermediate semantic web stored in Apache Jena.

Omega-PSIR data model

While designing the system and its functionality we had to consider such data structures that would be optimal for implementing planned functionalities, so that the needs of all user groups could be taken into account. Two main widespread approaches, namely VIVO ontology (Krafft et al, 2010), and CERIF (euroCRIS, 2016) were considered. Additionally, as a preliminary exercise we have also defined our own ontology (X3 et al 2013). Finally the accepted OMEGA-PSIR data model was dictated by practical solutions, resulting mainly from the simplicity requirements put on the model.

One of the main priorities was to implement the system in such a way that with limited programmer resources it would be possible to respond to changes in the system environment. To this end, one of our tasks was to invent a way to make the system flexible and responsive to changing requirements. Finally, we have decided to implement OMEGA-PSIR based on a NoSQL paradigm, though with a lightweight typing of data structures. We used XML means, namely XSD, for defining data model, which are stored in Apache Jackrabbit[4]. So within OMEGA-PSIR data are stored in the form of XML structures, compliant with a lightweight schema expressed in XDS.

Each entity type (*institution*, *person*, *book*, *paper*, *journal* etc.) has its own schema definition (in XSD form). For every entity it is possible to predefine various "digital attributes". The digital attributes are devoted to storing digital objects, which are then accessible through the OIDs. The text objects are subject to indexing, so that the index for full text retrieval is built automatically with new objects added to the database. The updates of text documents are automatically reflected in the indexes. In addition, with any defined entity instance in the knowledge database a unique identifier is assigned; this is non-changeable over the instance's life.

As entities are stored in XML form, each entity can be seen as a tree-like structure. Taking publication and an example, the tree root represents a single publication, and subtree nodes represent bibliographic elements, e.g. author(s), publications source, like books or journals. Each tree node might have its own properties (e.g. title, name, surname).

The entities may refer to each other, building a kind of semantic network of interlinked objects. The relationships between the entities are implemented by means of entity identifiers (logical link), optionally accompanied by some attributes of the related entities (entity embedding). In addition a relationship can be defined as *live* or *historical*. In the case when the relationship is life, the embedded substructure (local) is updated whenever the related entity (global) is updated. It may be useful for the cases, when the relationships should always provide the lastmost version of the related entity.

For historical embeddings, the changes of the related entity (global) do not affect the local embeddings. This approach makes possible to preserve historical version of the node in the local (sub)tree, and in the same time, link to the current version of that node. Separating global versions from the local ones makes possible to modify data on either side, e.g. change the author's affiliation only for a given publication (locally), or change authors name at a global level of the PERSON object, leaving the old name (historical) at the local author subtree within a given publication tree. In addition to preserving the historical values, this approach makes possible performing search in two ways:

1. text based search for author publications (just searching documents containing a given name string)

[4] Apache Jackrabbit is a schema-less solution, and is compliant with JCR repository standard (http://jackrabbit.apache.org/)

2. ontology driven search for author publications (whichever name is provided, all the publication of a given person are provided)

Mapping from OMEGA-PSIR to CERIF

As the data model of OMEGA-PSIR was optimized for the functionality and our needs, CERIF could not be used literally, it was rather used as a style. Nevertheless both models are very close to each other. Omega-PSIR model covers a very wide spectrum of research activities and results. It defines *inter alia* the following entity types:

1. *Person*
2. *Organisation* (used for providing affiliation of the person, but also as *corporate author*, *Publisher*)
3. *Publication* (various types bibliographically and contentwise)
4. *Journal* (with various attributes for quality evaluation, Sherpa-Romeo flags, etc.) , *conference*
5. *Grey Literature, including Thesis* (*PhD, MSc, BSc*), *technical report, presentation*
6. *Project* (accompanied with definable taxonomies
7. *Activity* (lasting in time) and *Achievement* (obtained in a given point of time)
8. *Product, technology, patent*
9. *Laboratory, equipment*

In addition a lot of auxiliary entity types are defined within the system.

Originally, the CERIF model is relational (ERM), but additionally there is a CERIF-XML format which has been agreed as the standard format for information retrieval from individual CRIS systems. The CERIF model is very broad and seemingly it is more extensive than Omega-PSIR. The current scope of the OpenAIRE information space, however, does not cover the full range of information in CERIF, therefore it is based on a subset of the full CERIF, namely covering *Publication*, *Product/Dataset*, *Person*, *Organisation*, *Project*, *Funding*, *Equipment*, *Service*.

While creating the full mapping between Omega-PSIR and CERIF would be eventually desirable, we focused on mapping between OMEGA-PSIR and the OpenAIRE subset of CERIF.

By now, we have tried three implementations of transformation:

1. On-the-fly conversion
2. Side-by-side conversion
3. Conversion using Intermediate RDF graph

Below we sketch the three approaches.

On-the-fly conversion

Omega-PSIR is capable of exposing its data *via* the OAI-PHM protocol and native services. OAI-PMH provides two xml formats for serving data: omegapsir and oai-dc. The latter one is obtained with the on-the-fly XSLT transformation, but it is also possible to achieve similar on-the-fly transformation with runtime-objects manipulations instead of xml manipulation. OmegaPSIR uses JSScript Engine for this purpose. It is natural to try such xslt/js approach for the transformation to CERIF, but the complexity of the mapping makes it inefficient.

Side-by-side conversion

To overcome the problem with slow performance of on-demand transformation from OmegaPSIR to CERIF, we can generate it once, and cache it for further requests. Clearly, the cache requires updating it periodically (or triggering by updates in the original database), but it fits well the characteristics of the systems, in the case there are a lot more reads than writes.

Conversion using Intermediate RDF graph

In this approach the full OmegaPSIR repository is transformed into one CERIF RDF graph stored in Apache Jena. The request for CERIF description of a single record are answered with simple SPARQL queries. The RDF graph generation has to be performed periodically. The big advantage, however, comes from the fact that this graph can be now used not only for interoperability with OpenAIRE. The SPARQL query language is more expressive that JCR XPATH, which is used in OmegaPSIR. By exposing a subset of OmegaPSIR as an RDF graph the access level to the Omega-PSIR reaches 5th level of Linked Open Data.

Literature

Bull, J., & Schultz, T. (2018). Harvesting the Academic Landscape: Streamlining the Ingestion of Professional Scholarship Metadata into the Institutional Repository. *Journal of Librarianship and Scholarly Communication*, 6(1).

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

Manghi, P., et al. (2012). The Data Model of the OpenAIRE Scientific Communication e-Infrastructure. Metadata and Semantics Research, *Communications in Computer and Information Science*, Springer, Vol. 343, pp. 168-180.

Príncipe, P., Rettberg, N., Rodrigues, E., Elbæk, M. K., Schirrwagen, J., Houssos, N., ... & Jörg, B. (2014). OpenAIRE guidelines: supporting interoperability for literature repositories, data archives and CRIS. Procedia Computer Science, 33, 92-94.

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

X1 et al (2015) "OMEGA-PSIR – A solution for implementing university research knowledge base." *EUNIS Journal of Higher Education* (2015).

X2 et al (2017). Integrating IR with CRIS–a novel researcher-centric approach. *Program*, 51(3).

X3 et al (2013). SYNAT system ontology: design patterns applied to modeling of scientific community, preliminary model evaluation. In *Intelligent Tools for Building a Scientific Information Platform* (pp. 323-340). Springer, Berlin, Heidelberg.