Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2014) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

CRIS 2014

RuCRIS: a pilot CERIF based system to aggregate heterogeneous data of Russian research projects

*Andrey Guskov[a], Oleg Zhizhimov[a], Vladimir Kikhtenko[b], Danil Skachkov[a], Denis Kosyakov[a]

[a]*Institute of computational technologies of the Siberian branch of Russian Academy of Sciences, 6 Acad. Lavrentjev avenue,Novosibirsk 630090, Russia*

[b]*Novosibirsk State University, 2 Pirogova Street,Novosibirsk 630090, Russia*

Abstract

The basic tasks of research administration are the registration and monitoring of projects execution. Currently, in Russia there are no centralized solutions for that purpose because of different requirements and regulations in national scientific foundations. The Ministry of Education and Science of Russian Federation initiated a pilot project for creating a prototype of a system for aggregating the information about research projects. The required system fits well with the CRIS model that harvests current heterogeneous research data from different sources, processes them and stores in a database, which finally provides aggregated information about national scientific research. In this paper some modelling, classifying and integration issues of this project are considered.

Peer-review under responsibility of euroCRIS.

Keywords: research project; CERIF; heterogeneous data; integration; classifying; XML; OAI-PMH

1. Introduction

The basic tasks of research administration are the registration and monitoring of projects execution. Currently, in Russia there are no centralized solutions for that purpose because of different requirements and regulations in national scientific foundations. The Ministry of Education and Science of Russian Federation initiated a pilot project

* *E-mail address:* guskov@ict.sbras.ru

Peer-review under responsibility of euroCRIS.

for creating a prototype of a system for aggregating the information about research projects. The required system fits well with the CRIS model that harvests current heterogeneous research data from different sources, processes them and stores in a database, which finally provides aggregated information about national scientific research. That is why we called it "ruCRIS". The similar projects having been developed in Sweden[1], Norway[2], Czech Republic[3] and some other countries for the years and have reached different stages of success.

Institute of Computational Technologies of SB RAS and Central Economics and Mathematics Institute of RAS executed this project in 2013. It includes the following main tasks:

1. Harvesting the data about research projects and their results from several preassigned sources.
2. Classifying, mapping and linking the data from various sources.
3. Storing the data in the central repository.
4. Querying and visualizing data from the repository.

The preassigned data sources were the databases and repositories of the following public foundations:

- Russian Foundation for Basic Research (RFBR),
- Russian Foundation for Humanities Research (RFHR),
- Center for Information Technologies and Systems for Executive Authorities (CIT),
- Republican Research Scientific and Consulting Centre for Expertise (RRSCC),
- Repository of Federal Programs of Russian Ministry of Education and Science (FP),
- Scientific projects repository of the Siberian Branch of the Russian Academy of Science (SBRAS).

2. Data heterogeneity issues

First, it was necessary to determine the list of harvested data entities. We identified main entities from project conditions using CERIF specification (Research project, Result product, Organization, Person), auxiliary entities (Contact, Description), link entities (e.g. relation between project and organization) and classifying entities.

Table 1. Data presence in external sources

Repository	Access type	Project	Result product	Organization	Person
RFBR	HTML	+			
RFHR	HTML	+			+
CIT	XML	+	+	+	
RRSCC	HTML	+		+	+
FP	XML	+		+	
SBRAS	DB	+	+	+	+

The first problem we met was syntactic and semantic heterogeneity of the original data in preassigned systems. As it is shown in table 1, all the data sources contain the information about scientific projects, but the rest of the data cannot be extracted in any appropriate way. Besides, there are three different ways of repository access, which require varied methods of interacting. First, when the data is published on the web site in human-readable way (HTML), it is necessary to use web-mining methods to collect it. Second, when the information is provided as files in XML format, it have to be processed by some kind of XSLT transformation. Finally, when the data source is actually a relational database, the information can be extracted by SQL requests. Noteworthy, the information obtained from HTML and XML (when XML schema is designed for other purposes) has poor quality and lack of attribute values. Having the database at the backend of this sources, it is preferably to develop the agent, which would interact directly with database and extract necessary information precisely for every single external source.

However, tight project deadlines and bureaucratic procedures did not allow us to get access to the sources in such way. Instead, we collected the raw data ("as is") from the available registries and then converted them to a common

format and data schema. We had to use different harvesting and transformation methods for different data sources: web mining, XSLT processing for XML documents, communicating by using OAI-PMH and SQL requests. We used the platform ZooSPACE (earlier – ZooPARK[4]) as integration bus that brings heterogeneous data to a single syntactic model. For each external data source, we developed the adapter, which converts the original dataset to the target format (CERIF). Thus, the result of the first-tier task was the array of XML documents containing data from various external sources in accordance with the scheme of CERIF-XML.

The second problem is classical for the master data management and system integration issues. It relates to the formation of a unified system for data classification and linking elements, which describes the same object in different datasets. There were allocated five main ways of master data management: centralized, translation table, consolidated, harmonized ("golden record") and transactional management[5]. For this task, the most appropriate way was using consolidated approach, when all the master data are managed in external systems independently. The central master-data repository tracks for any changes in external systems and then generalizes, normalizes, merges and purifies the master data. As a result, it forms the consolidated repository, which every record can contain one or more links to the source records in the external system.

According to this approach, the list of core classifiers was approved, which includes research areas, organization types, project result types, positions, academic degrees and so on. For every classifier we defined the basic set of classes and described the rules for mapping data from external source. During data processing, these rules were applied to every classifier entity, whereby the raw data proves to be reduced to a single classification scheme.

The linking data from different sources issue was solved for Persons and Organizations entities. Currently in Russia, there is no effective way to identify the person using some kind of personal UID. Therefore, there is no possibility to create the automatic procedure, which answers the question whether two records from different data sources contain the information about the same person or not. Thus, such a procedure could be built only using implicit features and should allow human-made decisions. Unfortunately, the raw data about persons contain insufficient information, which actually consists of full name, academic degree, affiliation and optionally something else. In this case, resolution must be made by matching these attributes and considering possible misprints and data changes (like affiliation or position change). Concerning organization identifications issues, there is an identification system in Russia (Taxpayer Identification Number), but it is not used in most sources. Therefore, for organizations we used the similar approach that also takes into consideration the different spelling of organization names (full ones or some kind of abbreviation). Based on this experience, we can conclude, that linking heterogeneous incomplete data requires for developing the hierarchy of rules, their weights and the conditions of applying them, that depend on the data sources completeness and their reliability.

3. Implementation

We used CERIF 1.5[6] for PostreSQL as a storage model. It allows us not only to use the unified XML-based format for data representation (CERIF-XML), but also to convert it into relational database model automatically. In this way, we have the following data processing conveyor (figure 1):

1. Data request.
2. Conversion into CERIF-XML.
3. Formal data check.
4. Data classification.
5. Data linking. Looking for the data that already exists in ruCRIS and linking it with the new records.
6. Data deduplication.
7. Storing data. Uploading CERIF-XML data into ruCRIS database using data access module implemented in ORM-style (Object-Relational Mapping).

The result of harvesting and processing heterogeneous data was the unified repository that contains the information about research projects, their executors and outputs (products). Actually, the quality of this data is not perfect, because they still contain duplicates and incorrect links. Nevertheless, it can be improved by more precise rules tuning and by adding other methods of quality control.

At the pilot stage of the project a web interface of the system was also developed. It visualizes the arrays of the research projects (cfProj), result products (cfResProd), persons (cfPers), organizations (cfOrgUnit) and their attributes, links and classifiers. It is notable that most of the sources provide the data only in Russian language, making them difficult to use in worldwide scope.

Usually, it is necessary to provide the possibility of "source linking" which establish relation between object in an integrated repository and origin source. For this purposes CERIF maintains "Federated Identifier" feature that allows specifying a number of identification services (cfSrv) for each origin source and multiple identifiers for every single object (cfFedId) used by different services.

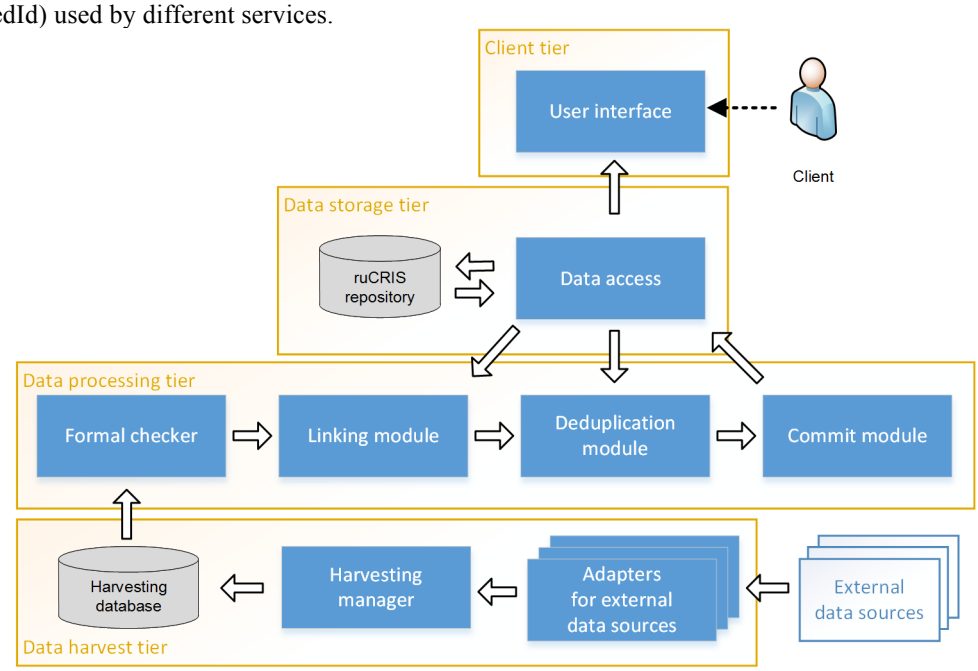

Figure 1. Architecture of ruCRIS

Finally, we have got the CRIS-CERIF repository, which contains more than 20 000 Russian research projects and about 10 000 interrelated organizations and persons. It shows that applied methods and technologies are suitable and promising for further development. If this project keeps running, one will get the possibility to obtain the answers on the complex questions about national research, such as «What research area was mostly funded in 2013?» or «Who are the expert/leaders in a research area X?».

4. Practical application and development

One of the conclusions obtained in this work was lack and inconsistency of data in the databases and repositories used. That is why it is so important to improve data collection from original sources – researchers and institutions. RuCRIS project is our effort to provide a kind of distributed platform for the data collection, analysis, interpretation and reporting.

Reporting part of ruCRIS project is of great importance to institution's public web sites construction. Our observations show that virtually none of Russian academic institutions public websites incorporates actual current research information. We suppose that such data is one of the most valuable parts of the institution public information profile.

Based on Microsoft Sharepoint cross-site publishing features we are developing a number of web sites, consuming major part of data from the common ruCRIS source. Main CRIS entities (organization units, persons, projects, publications, result products) are represented with corresponding web pages on the institution web site.

Such pages consist of web-parts that get their content from Sharepoint farm search index that is, in turn, populated from ruCRIS source. Web-parts are rendered asynchronously and obtain data in two different ways – one for interactive human-friendly representation and the other optimized for search bots. We also link full texts of publications in PDF format to CRIS metadata.

This approach allows us to represent common set of information (e.g. joint projects, publications) on different institutions web sites. Resulting web sites have a large number of pages, each with distinct URL representing particular entity in search-optimized form. Publications with full texts and/or abstracts are also indexed well by Google Scholar.

5. Conclusion

In conclusion, the CERIF model application and XML-representation allow us the effective implementation of the syntactic and semantic levels of heterogeneous data integration. However, it seems appropriate to propose euroCRIS to develop and propagate the recommendations, best practices and facilities for transport level of heterogeneous data integration. These include, for example, issues of communication protocols, such as REST-service facilities and protocol OAI-PMH[7] applied to CERIF data model.

References

1. 1. Johansson, Åke; Ottosson, Mats Ola: A national Current Research Information System for Sweden. In: Jeffery, Keith G; Dvořák, Jan (eds.): E-Infrastructures for Research and Innovation: Linking Information Systems to Improve Scientific Knowledge Production: Proceedings of the 11th International Conference on Current Research Information Systems (June 6-9, 2012, Prague, Czech Republic). Pp. 67-71. ISBN 978-80-86742-33-5.
2. Wenaas, Lars; Karlstrøm, Nina; Vatnan, Tore: From a national CRIS along the road to Green Open Access – and back again: Building infrastructure from CRIStin to Institutional Repositories in Norway. In: Jeffery, Keith G; Dvořák, Jan (eds.): E-Infrastructures for Research and Innovation: Linking Information Systems to Improve Scientific Knowledge Production: Proceedings of the 11th International Conference on Current Research Information Systems (June 6-9, 2012, Prague, Czech Republic). Pp. 289-294. ISBN 978-80-86742-33-5.
3. Chudlarský, Tomáš; Dvořák, Jan: A National CRIS Infrastructure as the Cornerstone of Transparency in the Research Domain. In: Jeffery, Keith G; Dvořák, Jan (eds.): E-Infrastructures for Research and Innovation: Linking Information Systems to Improve Scientific Knowledge Production: Proceedings of the 11th International Conference on Current Research Information Systems (June 6-9, 2012, Prague, Czech Republic). Pp. 9-17. ISBN 978-80-86742-33-5.
4. Review of Z39.50 servers and Z39.50 environment in Russia / V. Baranov, A. Plemnek, N. Sokolova et al. // Library Hi Tech.-2000.-Vol. 18.-N. 4.-p. 304-314.
5. Andryushkevich, Sergey; Guskov, Andrey: Practice of solving integration problems for information systems based on a master data management // Computational technologies. 2013. T.18. N 6. Pp 3-15 (С.К. Андрюшкевич, А.Е. Гуськов Практика решения задач интеграции информационных систем на основе управления мастер-данными // Вычислительные технологии. 2013. Т.18. № 6. С. 3-15).
6. CERIF 1.5 Reference. euroCRIS, 2013. http://www.eurocris.org/Uploads/Web%20pages/CERIF-1.5/cerif.html
7. The Open Archives Initiative Protocol for Metadata Harvesting. Protocol Version 2.0 of 2002-06-14. Document Version 2008-12-07T20:42:00Z. http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm