Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 33 (2014) 266 – 271

Procedia
Computer Science

CRIS 2014

Combining VIVO and Google Scholar data as sources for CERIF Linked Data: a case in the agricultural domain

Alberto Nogales[a], Miguel-Angel Sicilia[a], Brigitte Jörg[b]

[a] *University of Alcalá 28871 Alcalá de Henares (Madrid), Spain*
[b] *Jei Bee Ltd. London, UK*

Abstract

The needs of global science have fostered open access to the results and contextual information of research organizations at an international scale. This requires the use of standards or shared data models to exchange information preserving its semantics when transferred between systems. In that direction, standards as CERIF or projects as VIVO were developed to exchange or expose the scientific knowledge. Also, there are other sources of scientific information in the Web that are useful to complement institutional repositories and CRISes. The heterogeneity of data models behind each source in turn raises the need for mappings between them to ease interchange and aggregate information. In this paper, we present a tool that integrates three sources of research information and enables their aggregating and export into both VIVO and CERIF models. We present a case study in agriculture using OpenAGRIS, a bibliographic database linked to Web sources with more than 7 million records. Concretely, we describe the methods to combine Google Scholar data for the scholarly content indexed in OpenAGRIS and aggregating new information provided by the first one, using our tool. Finally the information is stored in a VIVO instance and then translated into CERIF using a conversion process mapping both data models. The case demonstrates the possibilities of mapping tools to aggregate and translate CRIS information.

Keywords: CRIS; VIVO; CERIF; OpenAGRIS; research information; Google Scholars

1877-0509 © 2014 Published by Elsevier B.V Open access under CC BY-NC-ND license.
Peer-review under responsibility of euroCRIS
doi:10.1016/j.procs.2014.06.042

1. Introduction

One of the objectives of researching is to share the knowledge with the rest of the world so it can take benefits of it. As most of the research projects are funded publicly, it makes sense to also have the results available for the rest of the world. It is normal to have access to papers, books or cases studies using the Internet so the rest of the scientific community or other people can apply it in their works. The amount of information is very big and there is also an interest in having an easy access to it. On that purpose of having the information available in heterogeneous formats some standards or projects have been developed.

One of the organizations creating a standard is euroCRIS[1], a not-for-profit association of Current Research Information System (CRIS) experts. A CRIS is a tool to provide access and to distribute scientific information. The aim of the organization is towards system interoperability, hence access and exchange mechanisms, guidelines or standards over scientific datasets or e.g. open institutional repositories. On that mission euroCRIS has taken the responsibility – as recommended by the European Commission – to continue the development of an interchanging format for data, namely CERIF [1]. CERIF is a formal data model for describing the scientific ecosystem e.g. how organizations are related and which are results of their works. The model was established as a storage model for the setup of CRISs and for the interchange of information between them, i.e. access to multiple, heterogeneous and distributed CRISs. CERIF was released 1991 and 2000 as an EC Recommendation to European Member States[2].

Another project with a similar aim of storing scientific knowledge and interchanging data between organizations is VIVO [2], providing an open source semantic web based application for the discovery of research information across institutions. It is based on an ontology through which institutions create their local instances and populate them with their research activities and results. Then the information is shared inline with the ontology enabling for wider discovery, networking and collaboration based on data about researchers and their works.

This paper presents a tool called agVIVO aimed to integrate three different sources of information in two different formats. The first source is a VIVO instance, which needs new information to be aggregated automatically. This information will then be used for retrieval with Google Scholar[3], the second source. The terms used in the Google Scholar searches will be published works on agriculture obtained from the third source, OpenAGRIS[4]. Finally the information will be formatted in two standards VIVO and CERIF.

2. Methods and materials

As we said before we are describing a tool and will apply it to a use case in the field of agriculture. Figure 1 summarizes the architecture of agVIVO, which is separated into two principle modules. The first module is called VIVO-io, which is responsible of aggregating new information for storage in a database to the VIVO instance. The second module is CERIF2VIVO whose aim consists of translating from VIVO format to CERIF and vice versa.

[1] http://www.eurocris.org/
[2] http://cordis.europa.eu/cerif/home.html
[3] http://scholar.google.com
[4] http://aims.fao.org/openagris

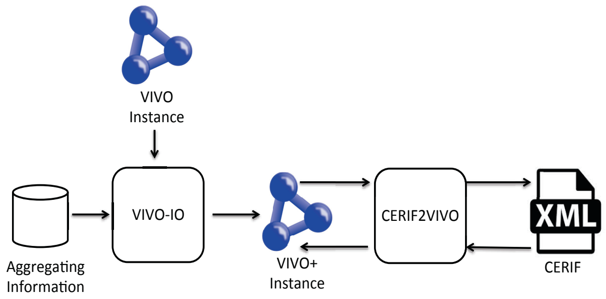


Fig. 1. Architecture of agVIVO.

2.1. Adding new information to a VIVO instance

Research organizations have the necessity of storing and sharing the information related to their projects. Work strategies, research bibliographies or reports to funders are essential outputs for other institutions or individual users. Making them public, the community will need to do smaller efforts in order to progress. As we have remarked before, the VIVO instance will contain research information stored by using semantic web techniques.

The first step in our approach consists of the automated aggregation of new data into a VIVO instance. A module developed in Java using Apache Jena[5], a framework to manipulate ontologies for accomplishing this task. The principle advantage of the proposal is that any information that can be represented inline with VIVO's ontology properties can be added automatically to the instance. The only problem is to have the information we are aggregating, stored at a source where to retrieve from. Using that module users avoid searching all the information they want and aggregate it to VIVO manually. They only need to find a source with the information and store it in a database.

2.2. Translating VIVO to CERIF and vice versa

VIVO is an approach to share research information but it has the problem that its use is not very extended, and that it is not recognized as a standard or recommended by any political community. On the other hand we have CERIF, which is a standard for representing research activities, entities including outputs. CERIF is a recommendation of the EU community and in used since the late 1990s. As CERIF will be continued by euroCRIS as a standard model for implementing a CRIS and exchanging information between systems, there is an advantage in translating VIVO instances into CERIF. Based on the mappings provided by [3] and [4], we have developed a translator between both formats. An example of a basic mapping between some of the principle terms in CERIF and VIVO is shown in table 1. Table 2 shows further mappings between CERIF and VIVO properties, though not as accurate as the ones presented in table 1.

[5] http://jena.apache.org/index.html

Table 1. Examples of mappings between CERIF and VIVO principle terms.

CERIF Table	VIVO Class
cfPers	foaf:Person
cfResPubl	bibo:Document
cfResPat	bibo:Patent
cfResProd	vivo:CaseStudy vivo:Dataset
cfFacil	vivo:Facility
cfSrv	vivo:Service

Table 2. Examples of mappings between CERIF and VIVO properties.

Table	Attribute	Class	Property
cfProj	cfURI	vivo:Project	vivo:webpage only vivo:URLLink
cfProj	cfAcro	vivo:Project	vivo:description only Literal
cfProj	cfStartDate	vivo:Project	vivo:dateTimeInterval only vivo:dateTimeInterval
cfProj	cfEndDate	vivo:Project	vivo:dateTimeInterval only vivo:dateTimeInterval
cfOrgUnit	cfAcro	foaf:Organization	vivo:abbreviation only Literal
cfOrgUnit	cfURI	foaf:Organization	vivo:webpage only vivo:URLLink

VIVO is represented through Resource Description Framework[6] (RDF) statements using classes and properties from the Web Ontology Language[7] (OWL). The CERIF exchange format is defined through the eXtensible Markup Language[8] (XML). We will apply the eXtensible Stylesheet Language[9] (XSL) allowing for the transformation of XML documents into other formats. Two stylesheets have been developed: one to transform VIVO into CERIF and another one to the reverse the transformation. Alongside the stylesheets a processor is needed for the conversion between both formats, where in our case Saxon[10] is being used.

The VIVO instance we are using is not important; probably any instance has a lack of information. The repository with the terms to find new information in Google Scholar is OpenAGRIS. This is a set of more than 7 million bibliographic references on agricultural research and technology and links to related data resources on the Web, like DBPedia[11]. We will use the titles of OpenAGRIS to search in Google Scholar new information like full-texts or references and will store it in a database. This information would be finally aggregated to the VIVO instance.

The way the information is added is inline with the following process. First we use the titles from OpenAGRIS to search new information in Google Scholar for storing it in a new database. Then we query all the titles from the VIVO instance within the database. If we find one occurrence that means we have new information to add. Hence we use the properties from the VIVO ontology for combining them with the corresponding values in the database, to add the information automatically. For example, we use the property "cites" to relate a paper with its references. Once we have added the new information we have a VIVO instance with aggregated data.

[6] http://www.w3.org/RDF/
[7] http://www.w3.org/TR/owl-features/
[8] http://www.w3.org/XML/
[9] http://www.w3.org/Style/XSL/
[10] http://saxon.sourceforge.net/
[11] http://dbpedia.org/About

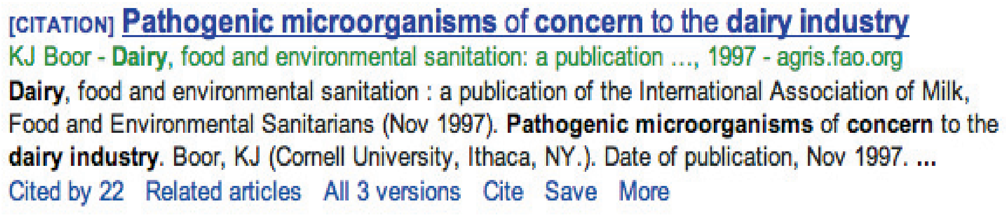
Finally the information stored in a VIVO instance is translated to CERIF. A benefit from translating VIVO into CERIF is that CERIF can be exposed as CERIF Linked Data.

3. Results and discussion

Using our tool we have been able to add new information to a VIVO instance and translate it to CERIF. Starting with the VIVO instance of Cornell University, we have found that e.g. the paper “Pathogenic microorganisms of concern to the dairy industry” written by Kathryn Jean Boor has no references. The following SPARQL query has been used to get the individual instances of the paper and then validate that it has no references.

```
PREFIX vivo: http://vivoweb.org/ontology/core#
SELECT ?subject
WHERE {?subject vivo:title ?title .
FILTER (REGEX(STR(?title), "title", "i"))}
```

This paper is also in OpenAGRIS. If we search it in Google Scholar we can obtain information that we are interested in but which is not included in VIVO, such as the references of the paper. In figure 2 a snippet of the paper including all paper-related information provided by Google Scholar is shown.



[CITATION] Pathogenic microorganisms of concern to the dairy industry
KJ Boor - Dairy, food and environmental sanitation: a publication ..., 1997 - agris.fao.org
Dairy, food and environmental sanitation : a publication of the International Association of Milk, Food and Environmental Sanitarians (Nov 1997). Pathogenic microorganisms of concern to the dairy industry. Boor, KJ (Cornell University, Ithaca, NY.). Date of publication, Nov 1997. ...

Fig. 2. Snippet from Google Scholars.

Through the module VIVO-io we can add references to the paper having it more comprehensive. The references have been obtained from Google Scholar and stored in a database. The code that is added is shown below in bold type. As we can see the property “cites” from VIVO ontology is used to relate a paper with its reference.

```
<rdf:Description rdf:about="http://vivo.iu.edu/individual/n58fi2cbppeiv5so3vacaa29eeg">
   <rdf:type rdf:resource="http://purl.org/ontology/bibo/Article"/>
   <vivo:title>Pathogenic microorganisms of concern to the dairy industry</vivo:title>
   <bibo:cites rdf:resource="http://vivo.iu.edu/individual/nu91cut1f9cgah0iio0graimm6q"/>
</rdf:Description>
<rdf:Description rdf:about="http://vivo.iu.edu/individual/nu91cut1f9cgah0iio0graimm6q">
    <vivo:title> Multistate outbreak of listeriosis</vivo:title>
    <rdf:type rdf:resource="http://purl.org/ontology/bibo/Document"/>
</rdf:Description>
```

Finally we can translate it into CERIF Linked Data, a new format for the storing the obtained knowledge about Cornell through a European standard not being used in the US. A research organization may be interested in working with the obtained results in some project.

In this paper we have developed a tool with the aim of solving the problems described above. Our development can aggregate new information automatically to a VIVO instance. Then this instance can be translated into CERIF format and in addition a translation between CERIF to VIVO can be automated. We have been able to combine three different sources of information: VIVO, OpenAGRIS and Google Scholar. The VIVO instance has finally also been populated with the information we are interested in. Finally we have presented the research information in two different formats.

Acknowledgements

The work presented in this paper has been part-funded by the European Commission under the ICT Policy Support Programme INFRA-2011-1.2.2-Data infrastructures for e-Science with project No. 283770 “agINFRA”.

References

1. Jörg, B.: CERIF: The common European research information format model. Data Science Journal. 9, 24-31, 2010.

2. Krafft, D.B., Cappadona, N.A., Caruso, B., Corson-Rikert, J., Devare, M., Lowe, B.J.: Vivo: Enabling national networking of scientists. In Proc. of the WebSci10: Extending the Frontiers of Society On-Line. , Raleigh, USA, 2010.

3. Lezcano, Jörg and Siciliar Lezcano, L., Jörg, B. & Sicilia, M.-Á. (2012). Modeling the Context of Scientific Information: Mapping VIVO and CERIF. In Proc. CAiSE Workshops (p./pp. 123-129)

4. Lezcano, L., Jörg, B., Lowe, B. & Corson-Rikert, J. (2013). Promoting International Interoperability of Research Information Systems: VIVO and CERIF. In Journal of Universal Computer Science, 19, 1854-1867.