DOI: 10.7250/eunis.2013.053

Innovative OpenSource Technologies for a CRIS: SURplus, a CINECA's solution

Federico Ferrario, Susanna Mornati, Davide Palena, Andrea Bollini, Sergio Bilello

Keywords
CRIS, Open Archive, DSpace, open-source, SOLR, Carrot2, search engine, Business Intelligence, SAIKU, datawarehouse, dissemination, governance, research, CINECA.

1. ABSTRACT

SURplus is the CINECA's best-of-breed solution for a CRIS, Current Research Information System. It is an IT platform that makes it easy to collect and manage data on research activities and outputs within an organization. Researchers, administrators and evaluators are given all the tools needed to monitor research results, enhance visibility and efficiently allocate available resources

One of the strengths of SURplus is that it sets new trends for IT support to CRIS, proposing and adopting innovative solutions to meet research need for governance. The decision to develop the solution on several open-source standards confirms this attitude. This paper describes the use of three solutions applied to different modules of SURplus CRIS platform.

Open Archive Module: DSpace

OA is the SURplus' module that manages collection and dissemination of research results. This module has been developed on open source software DSpace (dspace.org) as a customized version – *enterprise ready* – implemented and edited by CINECA. Upgrades of CINECA's *enterprise version* are periodically released to the open-source community as a contribution for software enhancement to reinforce CINECA's role and the investments of its clients. CINECA is also contributing to DSpace community as a Registered Service Provider and has an expert serving in the Committer Board. DSpace is the world's most widely adopted digital asset management system (923 out of 2,256 institutional and disciplinary repositories adopt DSpace, source OpenDOAR, Jan. 2013).

CINECA's Enterprise Version of DSpace pays particular attention to the simplification of data collection processes. It integrates into the platform a variety of services from commercial databases (such as SCOPUS and Web Of Science), free sources (as CrossRef, PubMed and ArXiv), reference management tools (as EndNote and RefWorks) and simple files (BibTex). Researchers can activate import from these sources or receive automatic notifications when a new publication with their name is found.

CINECA and Hong Kong University designed DSpace-CRIS, and CINECA released it as open-source. It allows the dissemination of description of entities in the research environment which go beyond publications: researchers, research units and organizations, projects and activities.

Expertise & Skills Module: Apache SOLR and Carrot2

DSpace-CRIS is the open source version of the SURplus' dissemination module called ES (Expertise and Skills), which allows the complete exposition of research activities and institutional assets and competences on the web.

The system permits free navigation of all the information through semantically rich interconnections among different entities (people, publications, projects, organization units, etc.) at a chosen level of granularity.

All data in OA and ES modules are also exposed through several protocols, standard formats as OAI-PMH, OpenSearch, RSS, CERIF, DC, MODS and interfaces like SOAP and REST to guarantee interoperability.

DOI: 10.7250/eunis.2013.053

The system has a very powerful search engine based on open source Apache SOLR (http://lucene.apache.org/solr/) that offers a scalable solution for indexing and searching millions of documents, offering valuable features like multilanguage search, synonyms search, faceted search, highlighting results, "more like this" hints and many others. All these features are carried out by means of Apache Lucene, another open source project of the Apache family on which SOLR is based. Apache Lucene is the effective search engine that is in charge of splitting data into tokens, extracting radixes (stemming), drop stop words and index them according to the analyzers chosen.

Every SOLR index can be customized adding specific metadata to accomplish particular search needs.

Moreover, SOLR offers a REST interface (XML , JSON), to allow loose coupling of applications that use search services and ensure their easy integration.

Carrot2 (http://project.carrot2.org/) is another open source project plugged in SURplus that allows runtime clustering of search results obtained through SOLR searches. In other words , Carrot2 Engine tries to infer categories from search results and classifies each search result in one of them. This is a very helpful feature since it organizes results in categories and makes it easy for users to spot the potential category of interest. Different algorithms are available (Lingo, KMeans, STC). All of them are based on Space Vector Models but each uses a different approach to find cluster labels, similarity coefficient, etc. Carrot2 clusters are displayed graphically in ES with Foam Trees or Circle.

A specific indexing of the information inside the Apache SOLR is used by the ES module to build up collaboration networks (co-authoring, co-investigation, etc.). Analyzing these networks brings to interesting semi-rough metrics (total number of collaborations, collaboration means for each researcher, variances, etc.) that can be elaborated within the Business Intelligence Module (BI) by comparing the metrics on different organizational levels of the Institution (work groups, departments, etc.) or by gathering them on other homogeneous criteria such as seniority, age, role, etc.

Business Intelligence Module: SAIKU

The third interesting open-source technology that is used within SURplus is Saiku (http://analytical-labs.com/), an open-source suite that allows user-friendly OLAP analysis of data marts within a data warehouse and, thanks to an interface based on front-end JQuery, it makes several operations available in drag&drop mode.

The separation between *front-end* and *back-end* in Saiku does not exclude the possibility to modify the MDX language behind the OLAP analysis, a language automatically generated by the *drag&drop* of the dimensions and the measures of the three *User Interface*'s sections: column, row and filter.

The Saiku's server follows the RESTful's methodology and it can interact with OLAP systems already in place, such as the MONDRIAN engine installed in the SURplus BI Module. Mondrian is an open-source ROLAP (relational online analytical processing) that translates the MDX query into SQL query based on the multi-dimensional model. This model permits to catch the results for optimizing the performances and it can be configured to restrict visibility.

Saiku has been chosen on Jpivot because it is more user-friendly and it makes it easier for the stakeholders to analyze data marts of a data warehouse. Besides that, the adoption of open-source solutions means reducing the license costs and it allows the SURplus team to customize and/or enhance the source code depending on the clients' needs.

Conclusions

This paper describes the use of three innovative open-source technologies applied to SURplus CRIS platform. DSpace, SOLR/Carrot2 and Saiku offer advanced solutions to submit, manage and retrieve data that are crucial to research governance in the context of CRIS applications.

DOI: 10.7250/eunis.2013.053

2. AUTHORS' BIOGRAPHIES

Federico Ferrario

Responsible for software development and architectures at CINECA since 2008. He has a long experience and skills in web architectures and J2EE platform. He has also developed an expertise in project management especially in university and government environments. Previous positions: Software Developer and Designer at CILEA (2001-2007). Education: Master Degree in Engineering at the Polytechnic of Milan, Postgraduate Master, II level, in Information & Communication Technology Management at the University of Milan Bicocca.

Susanna Mornati

Responsible for IT projects and services for research at CINECA since 2010. She holds a strong experience in project and program management in complex environments. She participates as an expert in the activities of several committees and serves on a number of international scientific boards. She writes and lectures on various aspects of IT management for research. Education: Master Degree in Linguistics at the University of Milan, Postgraduate Master, II level, in Information & Communication Technology Management at the University of Milan Bicocca.

Davide Palena

Responsible for software design and development of IT solutions for Current Research Information Systems at CINECA since 2007, with particular regard to interoperability and information dissemination. He has an extensive experience in design and set up of systems for research dissemination and managing critical issues in design of data models, business processes, system integration. He is also involved in the design and implementation of management systems for research evaluation and reporting. Education: Degree in Computer Science at the University of Milan Bicocca

Andrea Bollini

Responsible for software design and development of IT solutions for e-publishing and digital repositories services for research at CINECA since 2004. He has an extensive experience in setting up publications repositories and systems for research dissemination. Since 2007 he has been a committer of DSpace, the most widely adopted open-source software for digital repositories. Education: Master Degree in Applied Mathematics at the University of Rome La Sapienza, Postgraduate Master, II level, in Information & Communication Technology Management at the University of Milan Bicocca.

Sergio Bilello

Sergio Bilello is a software engineer who has worked at CINECA since 2011. In July 2012, he obtained a Master of Science Degree in Computer Engineering at the Polytechnic of Turin. He wrote his thesis on his job-based design and development of a Business Intelligence solution, realized with the Pentaho BI Suite. He is currently dealing with software development of IT solutions for Current Research Information Systems.