DEPLOYMENT OF A CURRENT RESEARCH INFORMATION SYSTEM (CRIS) FOR INTERNAL DISCLOSURE OF RESEARCH OUTPUT

Bart Goossens, Marc Pollet

Research Institute for Nature and Forest, Belgium
Bart.Goossens@inbo.be, Marc.Pollet@inbo.be

Abstract

In 2012, INBO decided to manage and disclose its research information by integrating its research output (papers, reports) with several other relevant data sources. To this purpose, an implementation project with Atira (Denmark) was started to deploy PURE as Current Research Information System (CRIS). This CRIS is based on the 'Common European Research Information Format (CERIF)' which allows a straightforward data exchange among European organizations. The system, however, also offers several useful services to INBO's own scientific and managerial levels. This paper describes the practical implementation of PURE at INBO, and its main characteristics.

Introduction

The research output of a scientific organization such as the Research Institute for Nature and Forest is diverse and encompasses scientific papers, projects, expertise and representations through networks and during scientific events. Next to that, it offers its researchers an array of external sources of scientific knowledge.

Until recently, separate systems were employed for each of those data sources. INBO's Information Centre adopted the Integrated Marine Information System (IMIS) to manage and disclose its own scientific papers as well as its journal and book collection. A procedure was operational for the deposition of each scientific paper by the researchers. Advisory reports are the output of a specific INBO business and the generation is supported by the INBO Advice Application (IAA). Metadata of projects, both administrative and scientific, are stored and managed in a Project Information System (PIS) based on JIRA. And people-related data is stored in the INBO People Application (IPA) and managed by the HR Department. Less attention has been drawn to information on the participation to networks and commissions, and to scientific events which was held in a local Excel file while information on the researcher's expertise was lacking all together.

This multitude of separate systems rendered a number of processes suboptimal, not in the least the time-consuming input of INBO's research output by the Information Centre's staff. Another critical drawback was the lack of an integrated disclosure of the above mentioned

information. In 2012, INBO looked out for a solution that tackled the above mentioned problems. In this process, we focused on implementing European standards such as CERIF in order to also enable an easy exchange of information internationally. For a number of reasons, both financial and functional, the application PURE (Publication and Research) of the Danish company Atira was selected.

In this paper, the data models supporting PURE are explained, the implementation project described and the major functionalities of this tool discussed.

Models

CRIS and CERIF

Since 1991 the Common European Research Information Format (CERIF) has been developed with the support of the European Commission and has been recommended to EU members for the storage and exchange of current research information (Jörg, 2009). In 2000, the custodianship of the CERIF standard has been transferred to EuroCRIS (Current Research Information System), a non-profit organization that maintains and further develops CERIF, and promotes its use.

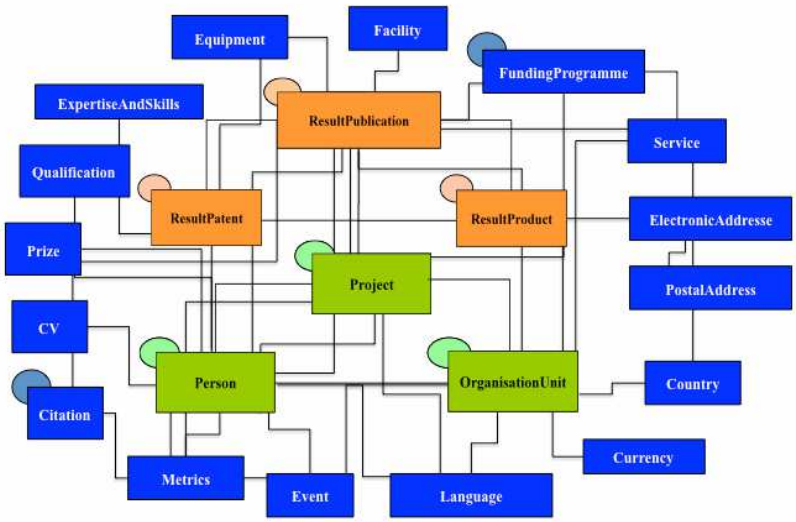

Fig. 1. CERIF entities and their relationships

Membership benefits of EuroCRIS are (i) professional contacts in the CRIS field and exchange of expertise, (ii) support in inquiries for expertise related to organizing and managing research information or building research information systems, (iii) the use of the

model for organizing scientific information, (iv) the opportunity to participate to Task Groups of one's interest, and (v) participation to conferences.

In recent years, the CERIF model has been implemented by some organizations for their own use. Other organizations developed a commercial CERIF-based CRIS such as PURE from Atira and Converis from Avedas (Russell, 2011).

CRIS activities and developments in Europe are tightly interrelated with CERIF. The physical CERIF model is a relational database model available as SQL scripts based on common ERM (Entity Relationship Model) constructs (Chen, 1976).

Research Information Systems (RIS) are built upon conceptual domain models to capture the meaning of the domain by structuring it into entities and their relationships (Wand & Weber 2002). The following entities make part of a traditional RIS: Person, Project, Organization, Publication, Patent, Product, Funding, Equipment, and Facility (see Fig. 1). An entity is represented by attributes and holds relationships with other entities. The relevant entities, their attributes and relationship descriptions as such, compose the model of the domain for setting up a particular information system (Jörg, 2009). In this business, *Current* Research Information Systems is preferably used to indicate their dynamics and timeliness (Jeffery & Asserson, 2006).

Atira and the open business model

Atira is a Danish software company, with a business domain in Research Information Management. It released PURE in 2003, now licensed for 47 900 research staff at 75 organizations (references) in 8 countries. (Atira, 2013)

Atira develops data models on three different levels, all based on CERIF: European, national and local (institutional). Local data models are based on the national one, and national data models are derived from the European one. Each of these data models is customized according to requirements specific of its level, although changes are usually minor. Costs related to the customization of a local data model are financed by the local institution, unless changes prove useful for the other organizations within the same country. In the latter case, costs are shared by the different organizations, or a central national organization. (Alroe, 2008)

At a national level, organizations that implement PURE are organized in a PURE User Community (PUC) that meets at regular intervals and discusses topics of mutual interest. This is also the forum to maintain and extend the national data model collaboratively. In Belgium, this PUC is moderated by the governmental agency Energy, Science and Innovation (EWI)

that also holds the licenses and maintenance contract (of all Belgian PURE users) with Atira. INBO is member of the Belgian PUC.

Belgian PURE data model

As mentioned before, any metadata model can be implemented in PURE: actually, an integral part of PURE's application architecture consists of a comprehensive technical framework. This allows a research institution to acquire PURE and still specify its own metadata model according to institutional needs and demands.

The High School of Ghent (HOGent) carried out the first implementation project of PURE in Belgium and consequently designed the Belgian data model. Of a larger number of entities, INBO uses the following: research output (publications), projects, activities, users, journals, publishers, organizations, persons, and events.

Implementation project and interfaces with other data systems

The PURE implementation project at INBO

Prior to an implementation project with Atira, a contract is generated based on requirements by the costumer gathered during earlier encounters. Both a (per day) price and planning are included and the project is started once the contract is signed by the costumer and returned to Atira.

The implementation project started with a two days start-up workshop at INBO. The purpose of this workshop was to discuss the major aspects and phases of the implementation project in order to make sure that both parties have a solid and common understanding of the project. During this workshop, PURE was selected as authoritative data source for some data, while authoritative data sources beyond PURE were identified. In the former case, a legacy import was worked out, in the latter a synchronization action.

The subsequent main part of the project was divided into iterations, each lasting two weeks exactly. Each iteration had a clear objective e.g. to get a specific part of the data integrations up and running. Each iteration started and ended with an online status meeting during which the past iteration was evaluated and the next one prepared.

A development server in Atira's secure hosting facility at Copenhagen was employed during development but the final application was moved to a final hosting facility in Belgium at the end of the project. Throughout the project INBO relied on a local PURE instance. INBO reviewed each iteration – for example parts of the data integration – in this private

environment. This offered the opportunity to reflect on data and functionality and to provide feedback.

Formal testing was done in two stages at the end of the project: an *Acceptance Test* confirmed that all functionalities were delivered as agreed upon by both parties in the contract, and a *Service Level* test confirmed that the system performed as agreed with full data loads, many simultaneous users, etc.

Integration with internal information systems

At INBO, PURE serves as integration platform for the institute's research output. As such, it would not replace extant businesses with an own management unit and possibly specific functionalities or e.g. administrative information that did not fit the PURE data model. In these cases, authoritative data sources were maintained but provide relevant data to PURE via regular synchronization actions (daily). Those data sources are:

- INBO People Application (IPA): data on employees of INBO (SQLserver database, Access front-end);
- Project Information System (PIS): project data (JIRA configuration);
- INBO Advice Application (IAA): data on advisory reports (SQLserver database, Access front-end).

On the other hand, as PURE only holds research output produced by INBO, it could not be used as repository for the (external) journal and book collection of the Information Centre. As a result, metadata on scientific publications were split into INBO papers that were imported into PURE (as a Legacy Import), and non-INBO publications that were further managed with Koha.

During a unique Legacy Import all INBO publication metadata were migrated from the IMIS application to PURE which thus takes over the exhibitor role of publicly available publication metadata records and possibly related full text documents. All imported content from the IMIS repository was properly related to other primary entities in PURE such as Persons and Organizational Units. Authors of publications were matched to the relevant Person records in PURE originated from IPA. Full person's names were synchronized into PURE. During name matching, we therefore used the last name of persons in PURE (from IPA) to retrieve a set of possible matches with the author's name. We screened the list of possible matches and created initials based on their first name. The initials created were compared with legacy data

initials. If one set of initials were found identical this was considered a match. If more than one set of initals were found identical, the match was considered unresolved.
Examples:
Van Waeyenberge, J. - Van Waeyenberge, John: match.
Stienen, E.W.M. - Stienen, Erhard Willhelm Max: match.
Van Waeyenberge, J. - Van Waeyenberge, Karen: no match.
Van Waeyenberge, J. - Van Waeyenberge, John and Van Waeyenberge, Jonas: no match.

All relevant data were added to a list with predefined fields (views) and Atira created a single run job based on the information we gave them to appoint the data from the view to the right fields into PURE. We also had to map the right views to the correct template types in PURE. The view PU1 holds all articles that appear in Web of Science, so they had to be mapped to template type 'A1: Web of Science article'.

The INBO PURE user application

The PURE application provides 5 tabs: Editor with the main data families (research output, projects, activities, etc.), Master data (holding lists of master data lists), Personal (with overview of personal achievements, and CV generator), Dashboard (not used yet), and Administrator (only accessible by the System administrator). Only one tab (Personal) is provided to the INBO employees, which contains all data families that are considered relevant (research output, projects, services, master theses, and activities).

Research output

At INBO research output holds the predefined templates "book", "contribution to book", "contribution to journal", "contribution to conference", "doctoral dissertation", etc. Each of those templates consists of template types that are created by INBO and correspond to the series or newsletters published by INBO (see Fig. 2).

Research outputs can be added manually or imported. Outputs are added manually by choosing the correct template. Users can change templates mid-submission if necessary, which avoids any unnecessary loss of data.

Full-text files are uploaded manually or imported when available in the source. The files are stored on the file server and can be set for available publicly, within the campus or not publicly available. Availability is set independently of the bibliographic metadata, which allows to be publicly available without the full text. The metadata in PURE is OAI (Open Access Initiative)-compliant. PURE incorporates both an OAI data providing mechanism and

an OAI data harvesting mechanism. The latter makes it possible to harvest metadata in bulk from external OAI sources, the former allows to harvest metadata by other systems from PURE. This is the case for INBO publications that will be harvested by DRIVER (Digital Repository Infrastructure Vision for European Research) Both the providing and harvesting mechanism of PURE provide metadata in Dublin Core format, CERIF-XML format and MODS.

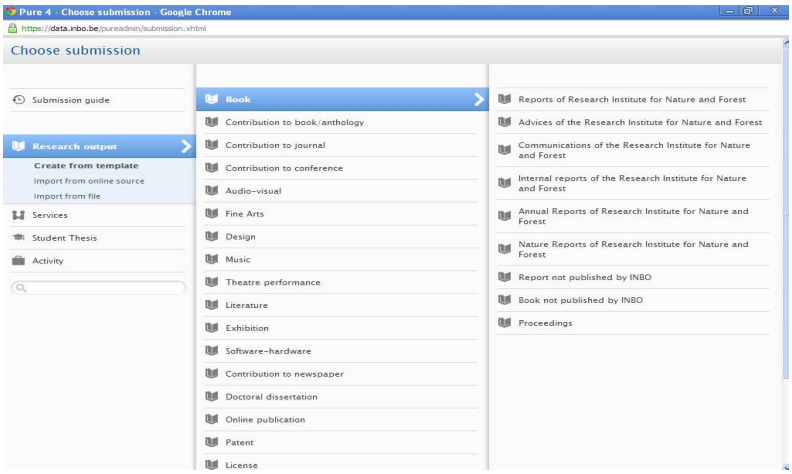

Fig. 2. Screen view of Research Output in PURE

Import of research output metadata can be carried out in three ways. A first type of import uses online sources like PubMed, Arxiv, Web of Science, etc. It must be taken into account that most of these sources require a license. Automatic author matching is part of this functionality; an algorithm analyses certain variables to make the right match between the authors in the imported record and the corresponding people's entree in PURE. Journals and publishers present in PURE are also matched, and if not yet in PURE, they are created automatically.

A second type of import can also be done from BibTex or RIS-files which are supported by a wide range of reference software (Reference Manager, Endnote, Refworks) and research databases (WoS, Scopus, etc).

A third way to a import output automatically is by author scanning. The scans are based on variations of the researcher's name and that of his organizational unit. This setting is made on the profile page of the researchers where variations of their name used per source can be modified.

PURE is using the Sherpa RoMEO API to display Open Access information about journals in PURE's user interface. The journal's RoMEO colour is shown in the publication template as soon as the journal is added. The user is encouraged to add the full-text. If the colour is GREEN, the full-text will then go directly online. If the colour is not green the researcher can set the correct embargo date (Atira, 2013).

PURE's workflow engine allows different roles. At INBO the researcher has the role of personal user and is able to create and send research output for approval. The Information Centre staff has the role of validator of these inputs. This makes it possible to enrich the data and to check the outputs on data quality. A research output at INBO is only validated if it contains a full-text document.

The Information Centre staff is responsible for the handling of (potential) duplicate publications. Facilities are available to specify the search by certain fields, by entity, by time periods, etc. The application offers both references from which the user selects one to delete. Relationships to other contents from the deleted ones will be moved to the remaining one. When merging duplicates, the user is shown the duplicates set in a special user interface where he or she can choose which parts of each duplicate will be retained - i.e. the author list from one duplicate, the abstract from another. Upon executing the merge, all relations to other content from all former duplicates are placed on the one unique publication that is the result of the merge.

Activities

Participating to a network or scientific event is registered in the entity 'activities' of PURE. Each researcher registers his own activities. The relationship to the event or organization (network) is made by the researcher. This information can be used both by the HR department (e.g. funding of activities), the management (e.g. evaluation procedures, governmental reporting, etc.) and the individual researcher (CV).

Master data

Master data are managed by administrators. An administrator can be made responsible for the entire PURE application or for a certain part (administrator of organizations, administrator of persons, etc.). Most of the master data like journals and publishers are managed by the Information Centre staff.

Classifications in PURE are fixed values usually displayed in a drop-down dialogue. The term "Classification" also applies to the concepts of controlled vocabularies and other types of taxonomies to be used in PURE. An example of a classification could be 'employment state', 'type of organization', 'publication state', etc.

Persons are not equivalent to users. For a person to become a user, at least one role must be assigned to him. Roles can be added and modified by Administrators. They can also lock a user account, create temporary guest accounts, and do other user management.

Reports

Reporting applies to every entity in PURE. Numbers and percentages of research outputs by types, discipline codes, internal authors and author rankings, peer review status, relations to events, relations to projects, relations to grants and funding bodies, publishing status, etc. can be generated.

Each report can contain multiple lists, tables, analysis results and graphs. Reports can be scheduled to run at a specified date and time and can be shared.

One of the major features of PURE is the integration of the metadata by creating relationships between entities (such as publications, activities, projects, or funding) which, obviously, allows extended reporting.

Personal overview

Personal is the researcher's main screen. It offers an overview of all personal content of the researcher including publications, activities, and projects. This page also provides overview of tasks, messages, favorites, curricula and recent actions, and gives access to PURE's online help system. The individual researcher can decide to make his profile publicly available, entirely or in part.

The profile itself holds information such as names and coordinates, name variations, titles, job functions, profile photos, scientific areas (controlled keywords; see also Classifications, above), attached documents, etc. Most of the information is already available from the integration with the HR system. Only non-synchronized fields can be enriched.

CV

In the personal section a researcher can create and manage his own CV which comprise content from PURE about the researcher’s publications, projects, activities, co-authors, etc. He only needs to add a section specify what content it should display – e.g. publications – and chooses a specific setting for that section. In addition to such content sections, researchers can also add free text sections and headlines in order to form a complete CV.

Sections with content will automatically be updated whenever new content (e.g. a new publication) is added to PURE. This ensures that CVs are always up to date. CVs can be published online on the PURE Portal, again, with updated content.

Further, it is possible to make certain restrictions/adjustments for sections with content. A section showing publications, for example, can be set to only show specific types such as

peer-reviewed articles. Sections can also be set to "Static" if the researcher wants the section to show particular outputs.

Researchers can view all of their relationships to other content graphically (this view is available throughout the application).

Dashboards

The Dashboard is an empty space where users can add Widgets as they like from a wide selection.

A widget displays a small graph or text-based piece of information as the result of a preprogrammed query on the PURE database. Examples are "Top ten cited researchers", "Top 10 most cited research outputs", "Activity types by year", etc.

Data access rights are controlled per users in the same way as reporting. Users cannot use the dashboard to access data they do not have appropriate rights to.

Finally, widgets, which hold limited resume-style information, offer the unique option of turning into a report at the click of a single button. The report can then be set to include more details about the specific area.

Administrator section

In this section the global administrator can handle most system management tasks which include: integration between systems, auditing log files, controlling sessions, managing all user's interface text, system messages and field labels, running special jobs; supporting users remotely, etc.

INBO website and PURE portal

All content in PURE, with "public" status is available from a built-in Web-Service API. The API makes it possible to disclose PURE content in the INBO website. Even other INBO web applications can retrieve data from the Web Service API, which makes it possible to use PURE as an integral component of the local Service Oriented Architecture. For sustainability reasons, however, INBO decided to harvest each separate authoritative data source through web services to feed its corporate website, including PURE for e.g. research output.

The PURE portal will be deployed operational for internal use only i.e. as tool to assign the right person (research expertise) to new research projects/outputs or research questions on the basis of his expertise.

Conclusions

The implementation of a CRIS at INBO enabled INBO to optimize and to simplify the processes for generation of research output. Both the researchers and library staff will no

longer use different tools for deposition, management and disclosure of INBO publications. Input has been facilitated through clearly defined template types and the import jobs from external sources.

The integrated disclosure of research output, advisory reports, project information, activities and personal data gives the researcher the opportunity to create a global overview of all his performances.

Pure provides an accurate, single source of information on the researcher's performances and the extended reporting facilities make it a very interesting tool for research managers to make decisions on future research projects.

Since 2008, the FRIS (Flanders Research Information Space) research portal (EWI, 2013) maps information about scientific research in Flanders. The CERIF standard allows INBO to respond to the commitments to transfer INBO-data to the Flanders Research Information Space.

The implementation project did reveal some technical issues that did not fit INBO's IT architecture. Data from three external data sources (IPA, PIS, IAA) are retrieved in PURE by synchronizations (ETL via views) instead of web services. The deployment of the tool on the tree extant environments (development, test, production) also proved suboptimal, e.g. because intermediate new versions of the tool cannot be skipped during upgrading. These shortcomings, however, are compensated by a number of advantages such as the PUC which can work as a permanent forum to address issues and change requests to the company and the use of JIRA as ticketing system. And Atira managed the implementation project in a very professional and committed way. Overall, INBO thus still firmly believes in PURE as corporate CRIS.

References

Alroe, B. 2008. The Danish PURE project: project model and system overview. Atira, Aalborg, Denmark.

Atira, 2013. [online]. Available: http://www.atira.dk [Accessed: April, 2013]

Chen, P.P. 1976. The entity-relationship model: Toward a unified view of data. ACM Transactions on Database Systems. 1, 1, 9-36.

EWI, 2013. [online]. Available: http://www.researchportal.be/en/index.html [Accessed: May, 2013]

Jeffery, K.G. & Asserson, A. 2006. CRIS: Central Relating Information System. In: Asserson A. and Simons, E. (eds). Enabling Interaction beyond the Hanseatic League. 8th International Conference on Current Research Information Systems, Bergen, Norway, 109-119.

Jörg, B. 2009. CERIF: Common European Research Information Format: Formal Contextual Relations to guide through the Maze of Research Information. In: Cvik, O. (ed). Proceedings of the International Conference CRIS Systems, Bratislava, Slovakia.

Russell, R. 2011. An introduction to CERIF. UKOLN, University of Bath, Bath, UK.
Wand, Y. & Weber, R. 2002. Research Commentary: Information Systems and Conceptual Modeling—A Research Agenda. Information Systems Research Journal. 13, 4, 363-376.