Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 106 (2017) 146 – 152

Procedia Computer Science

13th International Conference on Current Research Information Systems, CRIS2016, 9-11 June 2016, Scotland, UK

Integrating a national network of institutional repositories into the national/international research management ecosystem

João Mendes Moreira[a*], Cátia Laranjeira[a], José Carvalho[b], Fernando Ribeiro[a], Paulo Lopes[a], Paulo Graça[a]

[a]*FCT-FCCN, Lisboa, Portugal*
[b]*Universidade do Minho, Braga, Portugal*

Abstract

PTCRIS (Portuguese Current Research Information System - https://ptcris.pt/) is a program aiming at the creation and sustained development of a national integrated information ecosystem to support research management, according to the best international standards and practices. This ecosystem includes outcomes and outputs modules, in particular institutional repositories managed by the nationwide service RCAAP (Scientific Open Access Repository of Portugal – http://www.rcaap.pt/).
In order to achieve such vision, PTCRIS has two main goals. The first one is to define a regulatory framework based on the best international standards and practices. The second is to foster the adoption of such framework in the various information systems, both national (including RCAAP) and local (at the institution level).
This paper reports the context, strategy and work developed thus far to make the Portuguese repositories network (RCAAP), with more than 40 repositories, compliant with the national research ecosystem (PTCRIS) by adopting its regulatory framework. Integrating RCAAP within the PTCRIS ecosystem will greatly contribute to promote open scholarship, open data and open science.

Peer-review under responsibility of the Organizing Committee of CRIS2016

Keywords: Network of Institutional Repositories, Research management ecosystem, Regulatory framework, Open science

* Corresponding author. Tel.: +351 218 440 100; Fax: +351 218 472 167.
E-mail address: jmm@fccn.pt

Peer-review under responsibility of the Organizing Committee of CRIS2016
doi:10.1016/j.procs.2017.03.010

1. Introduction

This paper describes the role and importance of standards and guidelines on the integration of institutional repositories within broader ecosystems to better manage research, as well as, to better support open scholarship, open data and open science.

More specifically, it reports the context, strategy and work developed so far in order to make the Portuguese repositories network (RCAAP), with more than 40 repositories, compliant with the national research ecosystem (PTCRIS) by adopting its regulatory framework.

The first section describes the national research management ecosystem (PTCRIS), the national repository network (RCAAP) and also PTCRIS regulatory framework. The second section describes the challenges, requirements, scenarios and the methodology and tools used in order to quickly implement PTCRIS regulatory framework in the Portuguese repositories network. The third section will focus on the work and results achieved after the pilot.

2. Information systems description

This section describes the national research management ecosystem (PTCRIS), the national repository network (RCAAP), PTCRIS regulatory framework and ends describing the challenge of integrate RCAAP services, namely institutional repositories, into the Portuguese research management ecosystem PTCRIS.

2.1. PTCRIS

PTCRIS[(1)] (Portuguese Current Research Information System) is a program officially initiated in May 2014 by FCCN (Fundação para a Computação Científica Nacional), the FCT (Fundação para a Ciência e Tecnologia – the Portuguese Foundation for Science and Technology) unit responsible for planning, management and operation of the national research and education network,. PTCRIS main goal is to ensure the creation and sustained development of a national integrated information ecosystem to support research management, according to the best international standards and practices. To this end, PTCRIS aims to:

- Define **the regulatory framework** (https://ptcris.pt/quadro-normativo/) to be adopted by the various information systems;
- **Coordinate** FCT's systems integration in accordance with the standards framework;
- **Coordinate** integration of **external systems** (national and international) with FCT according to the standards framework;
- **Support** and **promote** the use of PTCRIS systems within the **community.**

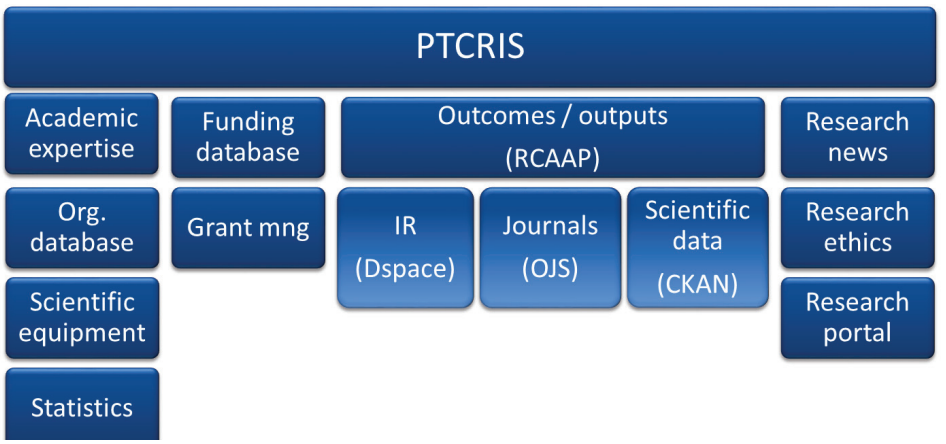

Fig. 1. PTCRIS components

2.2. RCAAP

A key module of the PTCRIS ecosystem is the outputs and outcomes component. In Portugal, this component is coordinated within the scope of RCAAP's[2] activity (Repositório Cientifico de Acesso Aberto de Portugal). RCAAP is the main instrument for Open Access (OA) implementation in Portugal. The initiative was launched in July 2008 and is currently carried out by FCT, together with the University of Minho. RCAAP offers a wide range of services in order to fulfil its OA mission. The most well-known electronic service is the hosting service of institutional repositories (SARI - http://www.projeto.rcaap.pt/index.php/lang-en/sobre-o-rcaap/servicos/sari) but there are other services provided by RCAAP, such as: hosted scientific journals (SARC - http://www.projeto.rcaap.pt/index.php/lang-en/sobre-o-rcaap/servicos/sarc), data repositories (SARDC - http://www.projeto.rcaap.pt/index.php/lang-en/sobre-o-rcaap/servicos/sardc), and a shared repository (http://www.projeto.rcaap.pt/index.php/lang-en/sobre-o-rcaap/servicos/repositorio-comum). OA resources are aggregated in the RCAAP Portal.

2.3. PTCRIS regulatory framework

To ensure that all components of the PTCRIS ecosystem work smoothly, a common regulatory framework had to be defined. Based on studies and state-of-the-art analysis the following standards and guidelines were adopted:

- **Data interoperability** - one of the main outcomes of PTCRIS is to achieve effective integration between elements of the ecosystem. To this end, both detailed application profile (CERIF-XML based), as underlying data format, and data semantics and classification are key;
- **Unique identifiers** – using unique, persistent and international IDs for researchers/teachers (ORCID), organizations (ISNI/Ringgold), etc. is paramount to break the information silos;
- **Data synchronization** - The ecosystem data currently resides in multiple, distributed national systems. In order to ensure data consistency, synchronization protocols must be used and PTCRISync[3] is one of the most used ones;
- **Privacy and Data Protection** - Building trustworthy indicators relies on having comprehensive, complete, consistent and reliable data. However, the need to collect, manage and use information may be at odds with data protection rights.

2.4. The challenge

This project involved three main challenges: The first challenge was to integrate RCAAP services, namely institutional repositories, into the Portuguese research management ecosystem PTCRIS. In order to achieve such purpose, two scenarios were considered: S1) adapt existing solutions; S2) change institutional repositories (IR) network to a software compliant with PTCRIS regulatory framework as much as possible. After some high level analysis it became clear that the most cost effective and sustainable solution was to move to DSpaceCRIS as this solution complies with many PTCRIS regulatory framework requirements, namely the fact that it is ORCID compliant[4] and it is a CERIF wrapped solution[5].

The second challenge was to take advantage of DSpaceCRIS, exploring functionalities that allow a better and a more comprehensive research management thereby providing advanced services to our member institutions. The DSpace-based repositories are aimed to be the place where researchers can deposit their scientific works but, in a broader spectrum, it lacks information about the researchers themselves, institutions or even funding programs. Such information gaps can be partially overcome by adding modules to DSpace to retrieve missing information. DSpaceCRIS is a DSpace fork, developed by CINECA, that allows for research information management beyond publications as it includes also information on researchers, funding, organization and other research entities. In fact, according to the survey report [6] carried out by EUNIS – EUROCRIS, CRIS-like repositories with extended data models are increasingly used as they provide a wide range of interoperability features between co-existing CRISs and repositories.

The third challenge was to respect the principles of open source, non-disruptive and sustainable approach. RCAAP IR network has more than 40 IR, being 28 of them managed in a SaaS (Software as a Service) regime. In order to test the approach, it was decided to run a pilot with three institutions: Universidade do Algarve, Egas Moniz and Universidade Aberta. This provid a safe environment to test the proposed approach.

3. Methodology

The adopted methodology for this pilot project has focused on the following points:

- **0 – Configuration and Setup** - Our infrastructure is virtualized which allows us to easily instantiate one server to create a new one or simply change DSpace configurations. To achieve automation we use a framework that permits a rapid deployment process, Rex, as the deployment and configuration management. This framework is written in PERL, easy to learn (some PERL background is needed), Apache 2.0 licensed and it's not intrusive as we don't need to have demons in our target machines. It uses ssh authentication, running a task directly in the machine. When there is at least one repetitive task that needs to run in multiple servers we need to automate it, here is when Rex comes into play. Basically Rex consists in a file (Rexfile) in which all tasks defined will be executed in one or multiple environments. It can be used to update the system or to deploy, compile and run code. Like any other PERL program we can also have modules that lie in the lib folder. The configuration files are written in yaml. As stated before, we created the necessary Rex tasks and configuration files of the machines that will be the pilot and execute Rex. The tasks included cloning of the DSpace CRIS code from CINECA, creating the database and compiling the code. The tasks can be executed independently which makes it possible to change some configurations or tweak some features individually or in multiple instances. Prior to instantiate all machines in our pilot, we created a development test machine to evaluate the process of installing and updating. Our pilot consists of 3 instances that need to be configured individually (layouts, Institution credentials, etc.) and have also common tasks. It would be an enormous effort if we had to manually install all the machines. Rex was used to get the code from the git repository, create and manage databases, change the specific configuration files of each machine (configurations defined in the yaml files), compile and then deploy, create users, create cron jobs and execute the code. The first challenge after was to make the master version of DSpace-CRIS, a fine tuned version able to replicate the application for the 3 instances. To this end, three main tasks were carried out: a) development of a new layout - By changing the homepage and some other aspects of the user interface; b) Translation - As all the applications we use are in Portuguese and English, we also provide a fully translated version of DSpace-CRIS; and c) Minor corrections - By testing the application we found some misconfigurations that we help to solve at the community level.
- **Kick-off meeting, basic configuration and Setup (1) -** A kick-off meeting with representatives of each institution was carried out to present the service, its objectives and align expectations. Then, each institution sent some basic information to allow the installation of each instance, namely the contacts of the management team institutional logos and information regarding the scope of the implementation (a department only or the institution as a whole).
- **Import of information (2)** - After the initial test, configurations and corrections, we defined the methodology of implementation along with the institutions. This consisted on importing existing or new information, then connect it and finally organize it on the application to provide a realistic view of the institution and the outputs. The underlying approach is depicted in Figure 2:

As the application manages different entities, the methodology has followed these entities as described: a) **Publications** (Including Thesis) - Publications were imported from the existing repository (DSpace) and, forone of the institutions in the pilot, from RCAAP's Common Repository (http://comum.rcaap.pt). Thesis were also included along with a specific identifier (TID); b) **Persons** - Persons are identified by the institution with the ORCID number

Fig. 2. Methodology of implementation

and a list is provided to be imported. If the name matches the author name already on the system, a connection will be made to associate this author with the ORCID account; c) **Projects** - The projects already exist integrated with an authority control functionality and will also be available with more information, a project page (as entity), of the projects of the institution; d) **Organizations** - The organizations will be identifiable in the future with ISNI numbers; and d) **Other** - For the pilot of DSpace-CRIS, other entities will also be exploited such as journals, events and datasets associated with the publications.

Once imported, information on publications, authors and projects will be associated. Then, the author/researcher can edit his profile with more accurate information, and the repository manager can associate the institutional affiliations.

To be able to login on the DSpace-CRIS instance, each institution created and ORCID APP to allow authentication methods.

A mandatory condition for the success of the association of the entities on the system is to have an updated list of authors with their author names (from the institution only) that match the author's name on the publication side. Then, the author itself can also update his profile with existing publications (from SCOPUS, or other ORCID data providers) and import them on the repository.

After the process of gathering all the information into DSpace-CRIS, the researchers selected by the institution to participate on the pilot will also step in and update their researcher page.

- **Improvements and new entities (3)** – Hereafter, the focus will be new entities like journals, events or datasets. These will be treated as independent entities, with more information to describe them. Consistency of the information will be tested by validating the information through the RCAAP Validator or through existing curation tasks of the DSpace-CRIS application.

- **Final Report (4)** - The pilot will be completed by July 2016, with a report from the institution gathering all the conclusions of the initiative.

4. Results

The provisional results of the DSpace-CRIS pilot can be categorized in different levels. On the technical part, we achieve some scalability regarding the management of different instances of the application. So if we need to create another instance we simply add the new server name to our servers.DEV.ini file and create a yaml file containing the configuration to the new machine. Then Rex will do the entire job.

This technical process easily allowed the creation of 3 instances based on the master version with the technical requisites of the PT-CRIS regulatory framework.

Regarding the application itself, the learning curve has not been difficult, as we had already experience in using DSpace, but some work has to be done to align the basic configuration with the needs of the project. As some core functionalities had not yet been finished, the import of information has to be made manually, which entails significant effort.

Aiming at developing a local CRIS, the selected institutions had already done extensive work regarding organization of the basic information which greatly facilitated its inclusion on DSpace-CRIS.

DSpace-CRIS allows a good connection between entities and provides an overview of the institution and the relations between entities. Also it allows the creation of local authorities that can be used to maintain an excellent quality of the metadata.

Some functionalities still need improvements, for example, the exposure of the CERIF-XML should be in a machine readable endpoint, instead of being present on the user interface. Also a mechanism to automatically sync the information, mainly for ORCID, should be developed.

By providing the same levels of interoperability and guidelines compliance, it allows to analyse the effectiveness of the application based on the real needs of the community.

5. Conclusions

IRs are an import part of the research information ecosystem. To smoothly integrate IRs into a ecosystem, they have to comply with a regulatory framework. This pilot has shown that, with the conversion of DSpace repositories into DSpaceCRIS repositories, it is possible to increase the level of compliance with PTCRIS regulatory framework, namely ORCID and PTCRISync compliancy. Additionally, according to CINECA, entity responsible for DSpaceCRIS development, DSpaceCRIS is expected to support OpenAIRE guidelines for CRIS systems based on CERIF-XML during the course of 2016. Thus, DSpaceCRIS is expected to deliver a high level of compliancy with PTCRIS regulatory framework.

On the other hand, over the last years, publications management systems like IR have become insufficient to address the needs of research information managers. Besides publications, information about authors, affiliations, funding and research infrastructures is becoming paramount. Though preliminary, our results suggest that with the right procedures and tools, it is possible to extend repositories into CRIS in a sustainable way.

This paper shows an approach to tackle both needs, i.e., to expand IR functions domain to CRIS functions and simultaneously integrate IR-CRIS in a broader ecosystem. By doing that we will contribute for a better support of Open Scholarship, Open Data, and Open Science.

References

1. Moreira, J. M. (2015). PT-CRIS Status Report, euroCRIS 2015, http://hdl.handle.net/11366/419.
2. Carvalho, José, João Mendes Moreira, and Ricardo Saraiva. "O RCAAP e a evolução do Acesso Aberto em Portugal." Uma Década de Acesso Aberto na UMinho e no Mundo (2013): 151-172, http://hdl.handle.net/1822/27919.
3. Mendes Moreira J, Cunha A and Macedo N. “An ORCID based synchronization framework for a national CRIS ecosystem”. F1000Research 2015, 4:181, (doi: 10.12688/f1000research.6499.1).

4. João Mendes Moreira, Paulo Graça, Bram Luyten, Andrea Bollini. “Developments with DSpace and ORCID Webinar Recording Available”, http://duraspace.org/articles/2511.
5. David T. Palmer, Andrea Bollini, Susanna Mornati, Michele Mennielli, DSpace-CRIS@HKU: Achieving Visibility with a CERIF Compliant Open Source System, Procedia Computer Science, Volume 33, 2014, Pages 118-123, ISSN 1877-0509, http://dx.doi.org/10.1016/j.procs.2014.06.019.
6. Ligia Ribeiro, Pablo de Castro, Michele Mennielli. “EUNIS – EUROCRIS JOINT SURVEY ON CRIS AND IR” – Final Report - http://www.eunis.org/wp-content/uploads/2016/03/cris-report-ED.pdf