3th International Conference on Current Research Information Systems, CRIS2016, 9-11 June 2016, Scotland, UK

PTCRIS_OrgID - Portuguese organisation identifiers authoritative system

Maria João Amante[a]*, João Duarte[b], Joaquim Godinho[b], Susana Lopes[a], Luís Quintano[b], Teresa Segurado[a]

aInstituto Universitário de Lisboa (ISCTE-IUL), Serviços de Informação e Documentação, Av. Das Forças Armadas, 1649-026 Lisboa, Portugal
b Universidade de Évora, Serviços de Informática, Largo Senhora da Natividade, 7000-810 Évora, Portugal

Abstract

Quality of data and interoperability are key for research information systems. Organisation identifiers enforce uniqueness and disambiguate institutional records ensuring necessary quality and consistency.

The main goals of the present work are to create rules, principles and processes regarding registration of organisations in Portugal, to build a first version of an authoritative National Organisations Database, to develop an organisation reconciliation service and to set up a registration system for those organisations.

The main outcome is the establishment of an organisation registration system and a mandatory national organisations database that syncs information between the main national and international registration systems like *Ringgold* and ISNI (ISNI+).

Keywords: Organisation Identifiers; Current Research Information Systems; CRIS; Interoperability

1. Introduction

FCT's (Fundação para a Ciência e a Tecnologia - the Portuguese Foundation for Science and Technology) goal is to establish Portugal as a global reference in science, technology and innovation, to ensure that knowledge generated

---

* Corresponding author. Tel.: +00351 210464052.
  E-mail address: maria.amante@iscte.pt

by scientific research is used fully towards economic growth and the wellbeing of all citizens. With this mission in mind, FCT started to work on making knowledge available to the public and industry by supporting open access to all scientific output that results from public funding. In order to reach the full potential of a national policy several tools need to be made available to all those involved. This paper aims to describe part of this national research information system specifically that which deals with organisations, firstly higher education organisations (public and private) and, later on, all organisations involved in the scientific system.

The scientific system has several actors/participants, from research institutions, publishers, industry, libraries, governments, financing agencies to researchers. Researchers are in the middle of all of this, producing knowledge through publications, patents, communications etc. and receiving funding from their organisation(s), funding agencies (national and international) and more. With this in mind, all monitoring schemes rely on input from researchers, from reporting and curriculum databases to open access repositories and complying with open access policies.

Researchers are asked to input their information on several platforms in many different ways and different moments of the scientific process, taking a lot of time and effort from their main task – science – and leading to noncompliance. To avoid this, we need to aim for an "input once, reuse often" principle. System interoperability is key to achieve this concept and connect all data through the supply chain making it easier to monitor and guarantee compliance.

To accomplish this, the PTCRIS program from FCT|FCCN (Fundação para a Computação Científica Nacional) aims 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.

2. PTCRIS

PTCRIS is a national Current Research Information System (CRIS) and its main goals are to define the regulatory framework to be adopted by the various systems, coordinate FCT's systems integration in accordance with the standards framework, and coordinate external systems integration with FCT (national and international) according to the same standards[1]. It is also important to support and promote within the community the use of PTCRIS systems.

In order to interconnect the different modules of PTCRIS it is important to use unique identifiers for all those involved and not only for individuals (ORCID). This means that organisations, research funders, publishers, industry and more, need to be correctly assigned to their product, may it be articles, funding/projects, researchers, etc.

2.1. Organisation identifiers (Org Ids)

What are unique identifiers for organisations and how are they used? Unique identifiers can be numeric or alphanumeric and are associated with a single entity. For organisations, an entity can be a university, faculty, department, research institute, company, etc. Identifiers are useful for disambiguating, consolidating and establishing a hierarchy view. They enforce uniqueness as they disambiguate institutional records, eradicate duplication of data, ensure correct delivery, entitlement and access rights, improve trust in data and map institutions into their hierarchy. By using unique identifiers for organisations, it is possible to link and synchronize different information systems, prevent duplicates, simplify data transmission and improve data quality.

Today we are faced with several data sources, data entered by different people and for different purposes, ending up with information that we cannot trust or validate. If we can minimize free-text forms and adopt increasingly the use of web forms, required fields, and web form consistency across systems and unique identifiers, we can improve data quality and cross matching between systems.

When using unique identifiers for organisations we ensure deduplication of records within and between systems, we enable seamless communication between organisations, remove ambiguity or lack of information, facilitating higher quality of data and better decision making for everyone involved.

2.2. International context: other projects (JISC, CASRAI, THOR)

Internationally, similar needs are under discussion by JISC and CASRAI (Consortia Advancing Standards in Research Administration Information) who, in 2013, produced a report on Organisation IDs, Landscape Study for

Jisc/CASRAI-UK Organisational ID Working Group[2]. The study examined the landscape of organisational identifiers in the UK and identified 23 different IDs and several use cases for organisational identifiers (Table 1). Use cases presented are related to authors/researchers, publishers, reporting and funding.

Table 1. Use cases identified.

Domain	Cases
Publishing	Research attribution Automatic allocation of Article Processing Charges (APCs)
Business Intelligence	BI for RPOs BI for research funders and regulators BI for publishers
Reporting	Reporting to research funders Reporting to regulators
Administration	Finance and billing Grant management
Sharing data/access control	Authentication and authorisation

This first study was followed in 2014 by a Review of selected organisational IDs[3] and development of use cases that aimed to investigate and review candidates for providing an authoritative, widely used unique identifier for organisations involved in research in the UK. The report concludes the best approach for organisation identifiers should be to use ISNI (International Standard Name Identifier) database in combination with the commercial option *Ringgold*, especially after *Ringgold* confirmed an ISNI number for every *Ringgold* ID and ISNI numbers assigned for all new records in the *Ringgold* Identify database. This solution is named ISNI+ and is illustrated in Fig. 1.
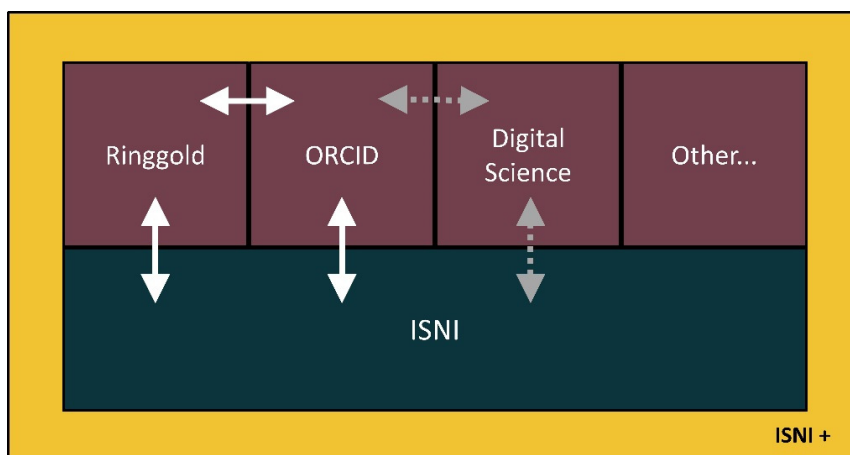

Fig. 1. ISNI+

Another project currently underway is THOR[4] funded by the European Commission under the Horizon 2020 programme to establish seamless integration between articles, data, and researchers across the research lifecycle. This project aims at interoperability between contributors, research output and organisations following work from project ODIN – ORCID (Open Researcher & Contributor ID Initiative) and DataCite Interoperability Network. ODIN was built on the ORCID and DataCite initiatives to uniquely identify scientists and data sets and connect this information across multiple services and infrastructures for scholarly communication. It addressed some of the problems identified

in the area like referencing a data object, tracking of use and re-use of data, links between a data object, subsets, articles, rights statements and every person involved in its life-cycle.

3. Our solution: PTCRIS_OrgID

Considering all of the above and under FCT's supervision, the Org-Ids project was created following recommendations of the study produced by the CASRAI-UK-Org-Id group. The major aim of the project is to prepare and define an action plan for adoption of unique identifiers for organisations.

The PTCRIS_OrgID project is ran by two teams, Universidade de Évora and Instituto Universitário de Lisboa (ISCTE-IUL) and started in October 2015. Each team is responsible for a set of tasks, Universidade de Évora for the technical developments and ISCTE-IUL for the rules and procedures associated.

The main goals were (task 1) to create rules, principles and processes regarding registration of organisations in Portugal, (task 2) to build a first version of an authoritative National Organisations Database, (task 3) to develop an organisation reconciliation service and (task 4) to set up a registration system for those organisations.

The main outcome was the establishment of an organisation registration system and a mandatory national organisations database that synchronizes information between the main national and international registration systems like *Ringgold* and ISNI.

First task was to carry out a prospective study of national and international agencies already involved in the registration of organisations, rules, standards and principles used. A survey was sent to national and international entities involved with organisation identifiers at a local or national level. Based on the results from this survey and the state-of-the-art study, procedures, standards and metadata were established for a National Organisations Database (PTOrg-DB) and an Organisation Registry System (PTOrg-IS) in Portugal.

CERIF (Common European Research Information Format)[5] was used as a widespread adopted format to achieve interoperability as it allows multiple representations of the same concept at different level of abstraction. The proposed metadata is CERIF/XML compliant so that the PTOrg-IS development follows the CERIF standards for data representation and systems integration.

Task two concerns building a first version of a *Ringgold*/ISNI compliant PTOrg-DB, based on national reference databases. The aim was to reconcile national registries, from FCT (research institutions) and DGEEC - General Directorate of Education and Science Statistics (higher education institutions), with *Ringgold*/ISNI registries. Some reconciliation services were analysed and conclusions of this study decided for the use of reconcile-csv (a simple csv reconciliation mechanism) with the OpenRefine interface.

The resultant PTORg-DB will be not only will be the first reference for the PTOrg-IS specification and development but also the reference for the reconciliation service that will be available for National institutions. This database is comprised of two-level hierarchy (organisations and sub-units) and is based in a CERIF-XML organisation profile to guarantee standard interoperability.

Such reconciliation service (task three) is also based on the reconcile-csv mechanism integrated with OpenRefine. The goal is that National institutions use such service so that, in the long term, all local repositories are "PTOrg-DB compliant" and so all organisation references are uniform: "Everyone speaks the same language".

PTOrg-IS full specification (task four) includes entities, metadata and functionalities description. The proposed base functionalities are:
- Organisation registry and management
- Organisation reconciliation
- Organisation search (public)
- Organisation sync with external *Ringgold* and/or ISNI databases

Following this work, such specification will lead to the implementation of a fully functional PTOrg-IS system (this implementation is not included in this specific project but has already been the target of an application for EU funds). Using as reference the "reconciled PTOrg-DB" arising from this work, the PTOrg-IS system will be the basis for all

the information systems of academic and R&D institutions. Additionally, the set of principles and rules defined will govern the several systems that include the PTCRIS ecosystem in alignment with the best international practices.

Future work will also aim at expanding the scope of the existing national organisational database to include more institutions, additional hierarchical levels and historical information.

4. PTOrg-IS Architecture

The system will provide an environment to create and maintain the Portuguese organisations with the following general goals:

- Fill a gap in our country as there is no reference system in this area;
- Enable each organisation to manage their own data (each organisation should designate one or more users for this purpose);
- Provide public access to data on the web;
- Provide access to data via an API (public and restricted environment);
- Guarantee the synchronisation with international reference databases;
- Configure multiple data access levels according to user profile,
- Guarantee that all the data is up to date;
- Avoid multiple records for the same organisation.

Four types of access profiles are defined:

- Administrator
- Organisation Manager
- Organisation
- Public

The Administrator profile grants full management of the system to a specific set of users. Those users can register, modify and delete organisations and are responsible to manage the system users. Managing the system, includes creating, editing, deleting and assigning profiles. An administrator can initiate a manual synchronisation and check the data of past synchronisations and has access to all data related to the modified organisation's history.

The Organisation Manager profile refers to a type of organisation. Users included in this profile can perform all administration tasks but only for their particular type of organisations. They are responsible for approving data change requests from organisations that are part of the defined type.

Organisation profile users are the owners of the organisation information in the PTOrg-IS. Those users may propose data changes in their organisation context, which has to be approved by the Organisation Manager user (according to the institution type). Users can also perform operations such as, search or export organisations.

The Public profile allows users to register an organisation. This type of request has to be approved by the Organisation Manager user. The users with this profile can query the data of organisations, perform searches and export filtered data.

All profiles are connected to each other, not only by sharing data (more or less restricted) but also through approval workflows and validation information; this will be detailed later in the document.

4.1. Approval workflow

Users can submit data to register, change or delete an organisation (classified with a specific type). An approval workflow, as shown in figure 2, is created and a task is triggered for the user who has the responsibility to approve the changes for the chosen organisation type (Organisation Manager).

If changes are approved, the organisation manager may complete or correct the submitted information and data is registered for internal update and synchronisation with external databases/systems. All users who participated in the process are notified.
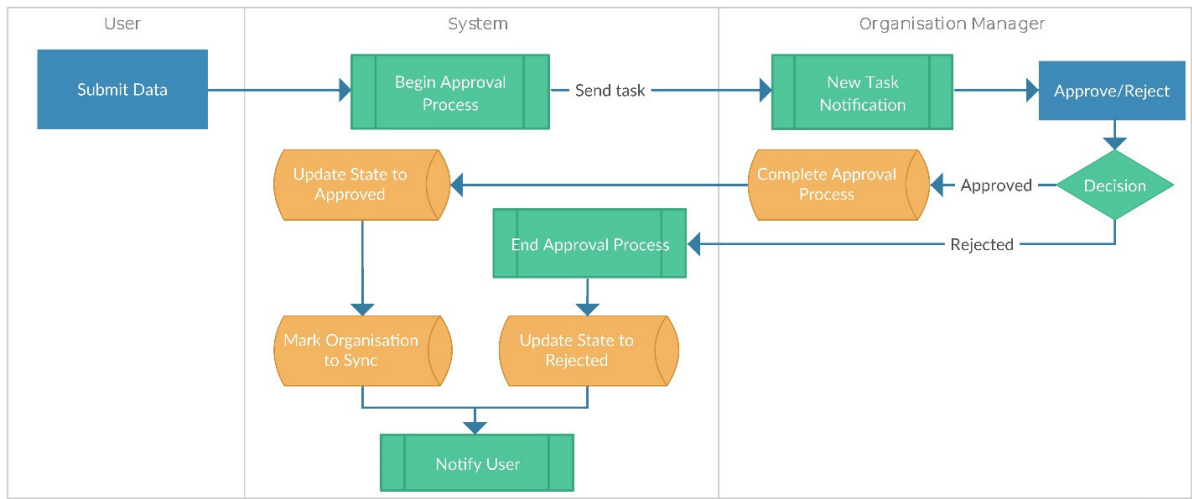

Fig. 2. Approval workflow

If changes are rejected, the workflow procedure terminates with a notification to the user that initiated the change proposal.

4.2. Database & Synchronisation

The database follows a CERIF implementation. CERIF (Common European Research Information Format) is a standard for managing and exchanging research data that includes information about organisations. All entities of the system are a subset or superset of the full CERIF model, developed and maintained by EUROCRIS.

The main entities of the system are:

cfOrganisationUnit (cfOrgUnit) represents the organisation,
cfPerson (cfPers), represents the users.

Other entities, like cfPostAddress (cfPAddr) or cfFederatedIdentifier (cfFedId) are used to store related information with the entities described above.

Another key feature is the automatic synchronisation with the ISNI+ databases and the integration with other systems that change the organisations' registries.

The following schema shows how the system (PTOrg-IS) interacts with the different systems (national and international as shown in the figure) in order to maintain his registries up to date and provide up to date information to the connected systems.
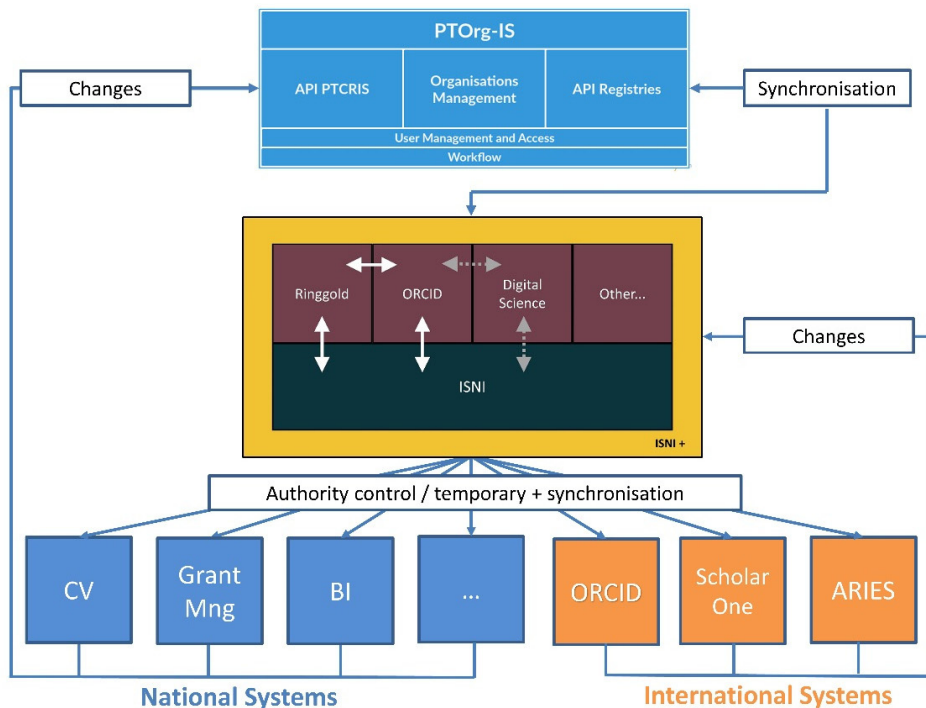
Fig. 3. Linked Data Environment

5. Conclusions

An integrated system such as the one outlined above will allow our country's organisations to have its information data normalised, up to date and available for public access. With API access the content can be created automatically or made widely available. The data is ready for easy sharing and redistribution.

The present work will support a national authoritative organisations database that will be part of the PTCRIS ecosystem as shown in the figure below.

This ecosystem aims to support scientific and research activity in Portugal.
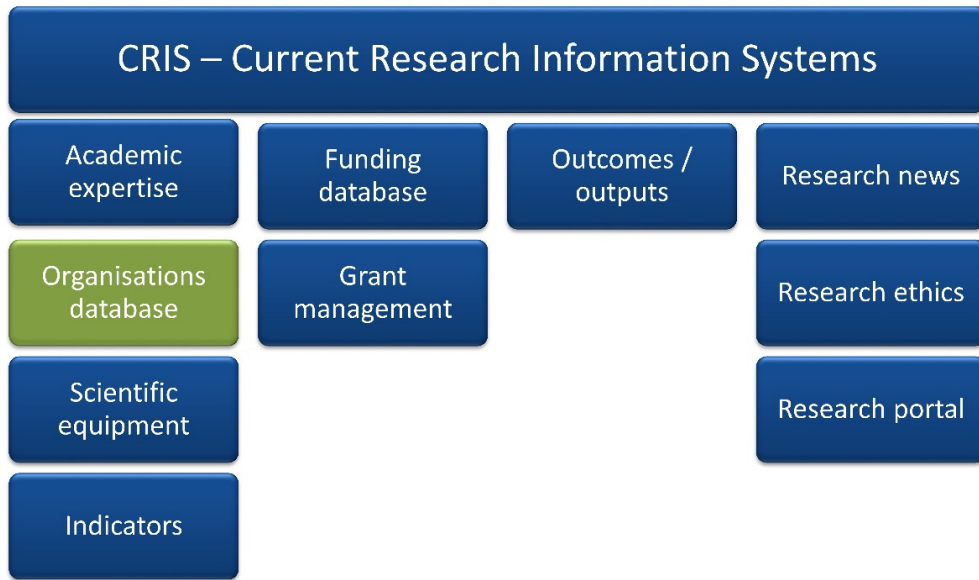
Fig. 4. PTCRIS Ecosystem

Acknowledgements

References

[1] Moreira JM, Cunha A, Macedo N. An ORCID based synchronization framework for a national CRIS ecosystem *F1000Research* 2015 **4**: 181.
[2] Hammond M, Curtis G. *Landscape study for Jisc/CASRAI-UK Organisational Identifiers Working Group;* 2013.
[3] Ferguson N, Moore M, Schmoller S. *Review of selected organisational IDs and development of use cases for the Jisc CASRAI-UK Organisational Identifiers Working Group;* 2015.
[4] Fenner M, Demeranville T, Kotarski R, Vision T, Rueda L, Dasler R, Haak L, Cruse P. *D2.1: Artefact, contributor, and organisation relationship data schema;* 2015.
[5] EuroCris, *CERIF 1.6 Model Info*; 2013.