Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2016) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

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

# Towards a CERIF-ORCID API adaptor: a progress report

Tom Demeranville[a]*, Josh Brown[a], Jan Dvořák[b,c], Dimitrios C. Karaiskos [b,d]

[a]*ORCID EU, Rue Dupré 15, 1090 Bruxelles, Belgique*
[b]*euroCRIS, Anna van Saksenlaan 51, 2593 Hague, Netherlands*
[c]*Charles University in Prague, Institute of Information Studies and Librarianship, U Kříže 8, CZ-15800 Praha 5, Czech Republic*
[d]*National Hellenic Research Foundation/National Documentation Center, 48 Vassileos Constantinou Ave., 11635 Athens, Greece*

**Abstract**

This paper describes an experimental implementation of the CERIF API specification being used as an intermediary system to enable the communication of information from the ORCID registry as CERIF XML. This work builds on the strategic partnership between euroCRIS and ORCID, and was enabled by the EC-funded THOR project. The implementation surfaced several issues stemming from the differing data models required, and from constraints specific to ORCID's position as an international linking infrastructure, such as privacy. The paper summarises the lessons learned from this endeavour, and sets out a clear roadmap and recommendations for the community to come together to implement, expand and sustain this potentially very valuable component of the research information ecosystem.


Peer-review under responsibility of the Organizing Committee of CRIS2016.

*Keywords:* ORCID; CERIF; ORCID API; CERIF REST API; Research Information; Interoperability

* Corresponding author. *E-mail address:* t.demeranville@orcid-eu.org

1877-0509 
Peer-review under responsibility of the Organizing Committee of CRIS2016.

## 1. Background

EuroCRIS and ORCID have had a strategic partnership since 2013. This partnership is based on the recognition that standards, like CERIF, and identifiers, like ORCID iDs, are essential to the success of large-scale, interoperable research information systems, at the institutional, national and international level. Without robust standards, the exchange and re-use of research information is fraught with difficulty.

CERIF CRIS systems do a lot to improve research information. They systematise, contextualise and enrich data from many sources, providing a rich and detailed view of individual and institutional research portfolios. Identifiers enable unambiguous connections between resources, as in the case of DOIs, or people, as in the case of ORCID iDs. ORCID is a community-driven, not-for-profit organisation which provides a free-to-use registry of unique identifiers for people involved in the research community. It was founded to solve the name ambiguity problem in scholarly communication, a goal that serves the CERIF CRIS community by improving the accuracy or attributions across the many systems that CRISs draw on to gather data. As of the time of writing, more than 2 Million people have registered for an ORCID iD.

ORCID is a partner in the THOR project, (http://project-thor.eu/) a European Commission funded project under the Horizon 2020 programme. THOR strives to establish seamless integration between articles, data, and researchers across the research lifecycle. This will create a wealth of open resources and foster a sustainable international e-infrastructure. The result will be reduced duplication, economies of scale, richer research services, and opportunities for innovation. The THOR vision is clearly analogous with the spirit of the euroCRIS and ORCID strategic partnership, and so when the opportunity arose within the project to explore ways of improving the interactions between CERIF-based systems and the ORCID registry, the team saw an excellent opportunity to generate an advanced case study of interoperability in action.

## 2. Requirements and rationale

A clear requirement emerged during the euroCRIS 2015 strategic members meeting to make it easier for ORCID and CRIS systems to interact, so that ORCID could become a more accessible information source for obtaining persistent identifiers and links from the wider academic communication world.

Current ORCID integrations within CRIS systems rely on the ORCID API. With this in mind, it was decided that the barrier to interoperability would be lowered if ORCID could utilise the same exchange format as the CRIS systems, CERIF XML. With the emergence of the CERIF API specification, there was also the timely opportunity for ORCID to support a standard protocol for exchanging CERIF XML. Implemented together, it was hoped that this would mean that CRIS systems could treat the ORCID registry in the same way as it treated its fellow CRIS systems. It would also make the ORCID registry the first public facing service to implement the CERIF API specification.

The core User Story can be stated as:

> *"As a client I would like to be able to retrieve CERIF XML from the ORCID registry for a person so that I can easily ingest and map it into my own systems."*

ORCID developed prototype support for the CERIF API and CERIF XML Entities with the technical support of euroCRIS.

## 3. Integration approach

In the first instance the API prototype was created as a read-only service separate from the standard ORCID API. This separate service implemented the CERIF API specification[1]. We decided not to overlay CERIF entity types onto the existing RESTFUL API using content negotiation for four reasons.

1. There was no clear way to map certain CERIF entities onto the ORCID API. For example, where ORCID has a single entity, "work", CERIF has two, "ResultProduct" and "ResultPublication". This could lead to a single REST resource serving multiple entities, which although possible, is not desirable.
2. The content types used by ORCID for content negotiation conflict with those defined by the CERIF API. The recommended content type for the CERIF API is "application/xml", which is an overloaded term within the ORCID API - it could refer to ORCID XML or CERIF XML. This can be overcome by defining a new content type for CERIF XML
3. Additional content types, especially at the prototype stage, cause maintenance overheads. Changes to the prototype code can affect the live deployed API so additional tests and procedures are required to modify it.
4. The two distinct entity mappings become intertwined to the point that they can be dependant on one another. This can cause problems if either model or API design change for CERIF or ORCID. This means API versioning becomes very difficult.

The open source CERIF API implementation was used as a reference and guide rather than integrated directly. The REST API and entity modeling was developed from the ground up in Java, building models from the XML schema using JAXB and serving them over a combination of the Spring and Jersey frameworks.

### *3.1 Semantic Layer*

For ease of implementation and due to time constraints, the implementation re-used the OpenAIRE CERIF semantic layer and classes[2]. It was felt that the majority of potential clients would understand and be able to utilise these semantics.

### *3.2 Entities*

The following entities were implemented: Person, ResultPublication, ResultProduct, OrganisationUnit and Funding. Email was deliberately omitted from the supported entities because the majority of emails within ORCID are private.

### *3.3 Privacy*

A number of issues arose around privacy and security while implementing the prototype. ORCID is a user centric system and supports three privacy levels, as decided by the ORCID record owners themselves. Privacy settings within ORCID are fine grained, with everything from name to email to individual works having their own privacy level. The levels are:

1. Public. Anything marked as public is displayed on the web page to all visitors and available via the Public API.
2. Private. Anything marked as private is only available to the ORCID record owner.
3. Limited. Anything marked as limited is shared with 'trusted parties' via the Member AP, i.e. client applications that the user has granted read-limited permission to during an OAuth sequence.

To replicate this level of functionality, it is necessary to apply OAuth authorisation to clients using the CERIF API. Although not described in the API specification, it is apparent that the API is authentication and authorisation neutral.

This means that although not a major issue, clients would need to understand the nature of OAuth and the OAuth application flow used by ORCID, increasing the barriers to interoperability.

In addition, ORCID XML, which has visibility attributes, does not map to CERIF XML. This meant that it is impossible to express which parts of an ORCID record the user considers limited access (i.e. for the receiving system only) and which parts they consider public.

## 4. Lessons Learnt

Once development was completed and reviewed, it was apparent that while it would meet the goals set out by the requirements, the prototype would not be suitable for production use.

1. More input from CRIS system vendors is required to ensure that the service fulfils their requirements.
2. Testing against real world systems wasn't possible. This is partly due to the early adoption nature of the CERIF API, but also because the end users of the ORCID CERIF API are still unclear.
3. There are concerns about expressing ORCID privacy levels within CERIF XML that need to be addressed.
4. We encountered problems with title languages when mapping works to publications and products. Language is a required attribute within CERIF XML, but generally unpopulated within the metadata held by ORCID. Simply defaulting to English is an option, but not a robust one. Another possible approach could rely on automated guessing of language.
5. Mapping from ORCID 'alternative names' to separate cfFirstNames and cfFamilyNames proved tricky. Although ORCID stores primary names as two components, additional names on a record are stored as literal strings. This was resolved using cfOtherNames.
6. CERIF API is authentication agnostic and does not specify an authentication layer. Our use case demonstrates that the CERIF API will need to tackle this, possibly by making its users aware that although authentication neutral, in real world deployments, authentication service layers (e.g. OAuth) could be seen and should be taken into account by client implementations.
7. The integration approach taken, while mitigating issues around diverging API design now and in the future, does not wholly remove the maintenance burden on the core ORCID registry. As a service owned by its members, it is difficult for ORCID to justify the required commitment at this time.

## 5. Future Work

ORCID remain committed to improving the links between the ORCID registry and CRIS systems but need to ensure that the provided solution is sustainable. We think the best way to achieve this would be to run it as a 'translation layer' or go-between service and talks are underway with euroCRIS and other partners with the goal of developing a shared open source CERIF-ORCID API adaptor tool. As well as enabling the code to be maintained separately from the core ORCID stack, it also allows us to form an open source 'community maintenance team' on Github with permission to merge pull requests into the repository. This would make the code much more amenable to changes in the CERIF space without impacting the core ORCID API. Likewise, as the ORCID API evolves we would be able to do our share and update the code. Taking this a step further, it could become a service that others host within their own infrastructure, configuring it to work with their API credentials.

Effectively it would be managed as an open source project - a partnership of ourselves, EuroCRIS and anyone else who is interested. Given the nature of the project, it is hoped that we will be able to engage CRIS system vendors in this effort so that they can shape, contribute to and enjoy the real world benefits it promises.

Exploring use cases further with CRIS users and vendors would be very worthwhile as emerging requirements could be met through other means, e.g. providing the ORCID data dump in CERIF XML format. By building on this

case study, we will be able to demonstrate concrete, measurable benefits from the strategic partnership between euroCRIS and ORCID, and provide further advantages for the research information community.

## References

1. Houssos, N., Karaiskos, D. (2015), CERIF REST API Specification v1.0, http://hdl.handle.net/11366/398.
2. Houssos, N., Joerg, B., Dvořák, J. (2015). OpenAIRE Guidelines for CRIS Managers 1.0. doi: 10.5281/zenodo.17065