Services that Enable Integration and Cross-Linking Across Different Types of Identifiers and Data Types

Document Information

Date: 13/09/2017

Authors: Guilherme de Mello (EMBL-EBI, orcid.org/0000-0002-9829-091X)
Markus Stocker (PANGAEA, orcid.org/0000-0001-5492-3212)
Angela Dappert (BL, orcid.org/0000-0003-2614-6676)
Robin Dasler (CERN, orcid.org/0000-0002-4695-7874)
Tom Demeranville (ORCID, orcid.org/0000-0003-0902-4386)
Kristian Garza (DataCite, orcid.org/0000-0003-3484-6875)
Florian Graef (EMBL-EBI, orcid.org/0000-0003-4890-5979)
Johanna McEntyre (EMBL-EBI, orcid.org/0000-0002-1611-6935)
Uwe Schindler (PANGAEA, orcid.org/0000-0002-1900-4162)

Reviewers Rachael Kotarski (British Library)
Ade Deane-Pratt (ORCID)
Laurel Haak (ORCID)

Abstract: This report summarises progress for disciplinary cross-linking of identifier systems and the results obtained from the perspective of each THOR project partner organisation, in particular disciplinary data repositories. We describe requirements, results, and challenges informed by implementations in the life sciences, earth and environmental sciences, and high-energy physics.

DOI 10.5281/zenodo.890959

This work was supported by the THOR Project. The THOR project is funded by the European Union under H2020-EINFRA-2014-2 (Grant Agreement number 654039). The following report is based on a deliverable submitted to the European Union on 30 May 2017.

Visit http://project-thor.eu for more information.

Executive Summary

In the modern research environment, researchers demand credit for the work that they do. While there are well established practices and services in place to give them credit for traditional publications, these are sorely lacking for the full range of research artefacts, including data and software. Researchers deserve credit for the time, energy, and expertise that they invest in creating, curating, documenting, and providing these research artefacts. Reliable attribution of credit requires unambiguous identification of the relevant components and links between them.

The THOR project supports the development of persistent identifier services, including ORCID identifiers for people and DataCite identifiers for research artefacts, as well as services that establish these links and connections among them. The approach pursued by THOR leverages global infrastructure services for persistent identifiers (PIDs), such as DataCite and ORCID, as well as more localised discipline-specific services that connect different types of data, to allow machine-to-machine communication through metadata.

This document describes efforts undertaken by the THOR partners to expand services that enable high-level cross-linking of systems across different types of identifiers. Each service provided by a THOR partner has its own use cases, requirements, and development considerations that are particular to the disciplinary and institutional context in which it resides. Overall, these services enable cross-linking across the different types of identifiers and data types that were most relevant in the context of that service, whether this be connecting data to publications, to researchers, to organisations, to funders, or some combination thereof. Each of these services builds important stakeholder support for the work of THOR and for Open Science among researchers, publishers, organisations, and funders.

This document presents the requirements for disciplinary cross-linking of identifier systems and the results obtained from the perspective of project partner organisation, in particular disciplinary data repositories.

Contents

1 Introduction

Today, researchers typically access scientific information via online resources, whether this be traditionally published resources, datasets, software, or even information about the researchers themselves. However, the lack of interoperability among the many available online information systems can make access more difficult than it needs to be.

The THOR Project has been working to improve the underlying interoperability to make this interconnected world of research possible by addressing issues of accessibility and discoverability. The approach pursued by THOR leverages global infrastructure services for persistent identifiers (PIDs), such as those provided by DataCite and ORCID, as well as discipline-specific services that connect different types of data. It also aims to enhance machine-to-machine communication.

In this document, we will describe efforts undertaken by the THOR partners to expand services that enable high-level cross-linking across different types of identifiers. Each service provided by a THOR partner has its own use cases, requirements, and development considerations that are particular to the disciplinary and institutional context in which it resides. Each of these services enables cross-linking across identifier and data types that were most relevant in its original context, whether this was connecting data to publications, researchers, organisations, funders, or some combination thereof. Each of these services builds important stakeholder support for the work of THOR and for Open Science among researchers, publishers, organisations, and funders.

This document presents the requirements for disciplinary cross-linking of identifier systems and the results obtained from the perspective of project partner organisations.

2 The Institutions and their Services

2.1 ORCID: Open Researcher and Contributor ID

ORCID is an effort to create and maintain a registry of unique researcher identifiers, and a transparent method of linking research activities and outputs to these identifiers. It is a hub that connects researchers with their research through embedding ORCID identifiers in key workflows, such as research profile maintenance, manuscript submission, grant application, and patent application.

ORCID provides two core services:

1. A registry to obtain a unique identifier and manage a record of activities
2. Application Programming Interfaces (APIs) that support system-to-system communication and authentication.

The ORCID registry is available free of charge to individuals, who can obtain and use an ORCID identifier, manage their record of activities, and search for others in the registry. Organisations may use a public API to collect ORCID iDs, and membership is required to request permissions from users to update ORCID records, and to receive updates from ORCID. Members and the community at large encourage employees, students, grantees, and so on, to register for and use an ORCID identifier. All data made public by its users are made freely available via periodic data dumps and the API.

The ORCID API allows systems and applications to connect to the ORCID Registry. This includes reading from and writing to ORCID records. The API is split into two parts: Public and Member. The Public API enables clients to read data marked as public by users. The Member API allows member organisations, who have agreed to the ORCID privacy policy, to request permissions from users to access non-public data and to write information to ORCID records. It also provides the ability to 'watch' ORCID records, and receive notifications when they are modified. Using API connections also provides transparency and authority, since it requires authentication and consent from the user, and also records provenance information.

2.2 EMBL-EBI: The European Bioinformatics Institute

The European Bioinformatics Institute (EMBL-EBI) is a centre for research and services in bioinformatics and part of the European Molecular Biology Laboratory (EMBL). Its core services include:

- ***ArrayExpress*** *– archive of gene expression experiments*
- ***BioModels Database*** *– a database of computational models relevant to the life sciences*
- ***Chemical Entities of Biological Interest (ChEBI)*** *– database and ontology of molecular entities*
- ***Ensembl project*** *– genome databases for vertebrates and other eukaryotic species (joint with Wellcome Trust Sanger Institute)*
- ***European Nucleotide Archive (ENA)*** *– resource of nucleotide sequencing information*
- ***Europe PMC*** *– database offering free access to collection of biomedical research literature*
- ***Experimental Factor Ontology (EFO)*** *– ontology of experimental variables for biomedical data*
- ***Expression Atlas*** *– database of summary information on which genes are expressed under which conditions*
- ***Gene ontology*** *– ontology of gene functions and processes*
- ***InterPro*** *– database of protein functional domains and families*
- ***MetaboLights*** *– a database for Metabolomics experiments and derived information*
- ***Protein Data Bank in Europe (PDB)*** *– European resource for the collection, organisation and dissemination of data on biological macromolecular structures*
- ***UniProt*** *– database of protein sequence and functional information (joint with Swiss Institute of Bioinformatics and Protein Information Resource)*

All these databases cross-link to each other and to the scientific literature, providing a deeply integrated ecosystem that reflects the natural connectivity of the life sciences. The cross-links are added by both data submitters and curators, depending on the type of database. The EMBL-EBI resources assign their own identifiers to their data records, although some groups assign DOIs to high-level datasets, for example database releases. These assignments are considered as an additional, complementary identifier to the experiment, rather than a data-record level identifier.

Given that these resource-specific identifiers are embedded in the practices of this scientific community, aligning them with more universal PIDs for data will be critical in emerging solutions, such as Scholix, in order to give a complete picture of life science data literature-data integration. This is particularly important as these databases have a high rate of reuse.

2.3 PANGAEA: Data Publisher for Earth & Environmental Science

PANGAEA, the Data Publisher for Earth & Environmental Science, is an information system that operates as an open-access library aimed at archiving, publishing, and distributing geo-referenced data from earth system research. PANGAEA is open to any project, institution, or individual scientist who wants to use, archive, or publish data.

Scientific data and related metadata are archived in a relational database. Published data are freely available and are distributed online in standard formats using Web services. Data are identified and citable by DataCite DOIs, and can be published as supplements to scientific articles or as collections in journals. Retrieval of data is supported by a full-text search engine and faceted search.

The PANGAEA data editorial process ensures the integrity, authenticity, and high usability of the data published. PANGAEA guarantees the long-term availability of its content through commitment of its hosting institutions, the MARUM Center for Marine Environmental Sciences and the Alfred Wegener Institute.

At PANGAEA, cross-linking identifier systems have a fairly long history. This history is arguably best exemplified by the data-literature cross-linking service PANGAEA has been operating in collaboration with Elsevier since 2010. This integration has enabled the cross-linking of data published by PANGAEA and articles published by Elsevier, and improved Elsevier ScienceDirect user experience by supporting the inclusion of geospatial widgets that visually show the location of supplementary data published by PANGAEA, and directly link to data. Vice versa, PANGAEA provides users with information about the articles that are supplemented by the published data. Such cross-linking provides users with contextual information and supports credit for publishing data. This integration is now gradually being superseded by Scholix.

2.4 CERN: European Organization for Nuclear Research

The CERN Scientific Information Service (CERN-SIS) maintains several information resources relevant to different stages of the research lifecycle for researchers in High Energy Physics (HEP). There is already a long-standing history of cross-linking in Inspire, the HEP literature aggregator that is the flagship resource of CERN-SIS. Within Inspire, multiple realisations of a single concept of a paper are united into one record (for example, the arXiv preprint of a paper and its final published version from a journal). Inspire typifies the CERN approach to information resources, seeing the automation of processes as a value-added service that spares the effort of researchers and encourages their continued use of the overall information resource.

Under the umbrella of THOR, this long-standing approach to centralising the collection of information and adding a layer of valuable automation is being extended to the newest information resource being developed by CERN-SIS, known as CERN Analysis Preservation. This is an internal CERN tool that seeks to link the disparate components that make up a conceptual HEP analysis in one place. In HEP, an ‘analysis’ is generally the set of data, software, internal notes, and other materials that comprise a particular investigation. The nature of this investigation and the nuances of an analysis in practice vary in definition among the experimental groups. CERN Analysis Preservation must be flexible enough to cater to the particular needs of these different experimental groups while still providing consistent added value for all.

2.5 DataCite: Leading Global Provider of DOIs for Research Data

DataCite is a global non-profit organisation that provides persistent identifiers in the form of DOIs for research data and other research outputs. DataCite works with its members, who allocate DOIs. These members enable data owners, stewards, or archives to assign persistent identifiers to research data. Members also encourage the use of best practices in research object citation, fostering cooperation with other organisations and entities, and provide mechanisms for community input and involvement.

DataCite provides a myriad of services to support the search and discovery of research data:

- DataCite Search provides an integrated interface, where it is possible to search, filter and extract details from a collection of millions of DOI records.
- DataCite Event Data retrieves and exposes the activity that occurs around research datasets. In particular, Event Data brings to light links between data and publications, software repositories, and people.
- DataCite Content Resolution exposes the metadata stored in the DataCite Metadata Store using multiple formats. It can also redirect to content hosted by DataCite participating data centres, allowing data to be accessed directly using a DOI.

Providing services that enable integration and cross-linking across different types of identifiers has been an activity into which DataCite has put significant effort and will look to improve in the future.

2.6 The British Library

The British Library (BL) is the national library of the United Kingdom and one of the world's greatest research libraries. It supports the UK's research infrastructure, serving business and industry, researchers, academics and students, world-wide as well as in the UK. The British Library is committed to improving access to, reuse of and tracking impact of born-digital material, including websites.

Persistent identifiers are a key technology in assuring access and preservation to the BL's collections. The British Library is a founding member of DataCite, which assigns persistent digital object identifiers (DOI names) to research datasets, and it is the UK DataCite registration agency; it works with over eighty of the leading data centres, research institutes, and research universities across the UK. The BL also supports the International Standard Serial Number (ISSN) UK Centre, which is responsible for assigning ISSN to serials published in the UK. The ISSN is the international standardised code that identifies all serials, journals, magazines, and periodicals, irrespective of their medium.

The International Standard Name Identifier (ISNI) is an International Organization for Standardization (ISO) standard number for identifying contributors to creative works and those active in their distribution, including researchers, inventors, writers, artists, visual creators, performers, producers, publishers, aggregators, and more. It is part of a family of international standard identifiers that includes identifiers of works, recordings, products, and rights holders in all repertoires. The British Library represents the Conference of European National Librarians on the governing Board of the International Agency (IA) for ISNI. The BL also provides quality assurance services and expertise in support of the ISNI assignment system and database. The BL has contributed to the ISNI database data from its EThOS electronic thesis service, its Electronic Table of Content service (covering 20+ years of articles from the top 20,000 academic journals in its collections), and c.45,000 identities created as part of the JISC Names project. The ISNI-IA, which was established in 2010, developed its initial database in 2011; it now houses c. 8 million identities. This deliverable discusses the development of a linking service between ISNI and ORCID person identifiers.

3 Requirements for Integration and Cross-Linking

THOR partners are working to deliver services that enable high-level cross-linking across different types of identifiers. These services are specific to the needs of the data archives in which they are implemented and work towards the common mission of identifying data provided by distinct researchers from a specific organisation and supported by a specific funder. These services provide essential stakeholder support for publishers, research organisations, funders, and the individual researcher in the form of services specific to the communities in which the THOR partners are embedded.

Because services exist within the particular ecosystem in which they are developed, this work took on a number of disparate subtasks, all under the umbrella of contributing to the global cross-linking mission. The THOR partners were variously charged with:

- Providing unified cross-linking services in their respective data systems to identify relationships between the data entities and associated metadata using PIDs
- Providing services to assign DOIs to datasets that have existing identifier systems and work with publishers to synergise artefact-contributor-organisation workflows
- Implementing prototype services to link works claimed retrospectively into ORCID records to both funding and affiliations
- Improving the person identifier linking service across ORCID and ISNI

The work of the THOR partners has achieved these goals, resulting in service improvements that will benefit the specific needs of each partner's immediate stakeholders as well as other organisations in the data services space seeking to implement similar improvements.

3.1 ORCID

ORCID's mission is focused on providing the community with links between people, places, and things. The THOR project's requirements for work from ORCID are therefore naturally orientated towards that mission. ORCID develops requirements for its services and metadata in partnership with the community that it serves, through working groups. Both prototypes and core functionality built through the THOR project are developed with these community needs in mind. Sometimes this takes the form of responding to community demand; other times it means building prototype services to show what is possible and passing the feedback to ORCID to include in their community discussions.

ORCID recently initiated a facilities working group[1], which will examine the requirements that large research infrastructures (RIs) have around linking people to funding, scientific facilities, organisations and research objects produced using these facilities. This group consists of representatives from RIs, publishers and funders, and will meet five or six times during 2017. They will discuss the requirements as they exist now, define the idealised workflows, and reach out to platform vendors to encourage adoption of requirements.

When considering linking works to funding and organisations within ORCID, there are two important points to consider. First, although ORCID can facilitate and represent these links, it is primarily concerned with the links between *people* and their works, affiliations, and funding. Links between funding and the works supported by that funding can usually only be inferred from their common connection to the

[1] https://orcid.org/about/community

individual owner of the ORCID record. Second, affiliation and work metadata on a given ORCID record are *part of that record* and should not be addressed individually, except as part of regular API interaction. This means that it is not possible to say work #3 on a record is linked to affiliation #4, unless they have matching external identifiers. Third party services can use the information in the ORCID registry to establish connections between works, funding, affiliation, and so on. This is usually undertaken in reporting systems or Current Research Information System (CRIS) implementations, for example.

There are several good reasons for this modelling beyond ORCID's person-centric conceptualisation of the scholarly record. Linking a work within an ORCID record to funding should not depend on that funding being referenced elsewhere in the researcher's activities. In addition, the availability of external persistent identifiers is better exploited by referencing them directly rather than indirectly. Users of ORCID metadata would then not have to make multiple requests to discover the funding of a work; rather, that information would be encapsulated within the work itself.

Several other requirements have emerged from project partners and research undertaken by the THOR project, such as modifications to the ORCID API that include search improvements, interface improvements and minor schema changes. These service improvements are described below.

3.1.1 Prototype Funding-Work Link Tool

ORCID EU developed a prototype service to address the requirement for a researcher to link existing papers with the funding that supports them. The end goal is to implement efficient workflows to collect IDs and connect people-funding-works-affiliations as the works are created or published, so that retrospective claiming is no longer required. It will take considerable time before this is widespread and even then, some combination of API and active linking may continue to be necessary. The prototype provides a model for an active service. The prototype developed, and the method it uses to create links between works and funding, is based on the use of external identifiers, such as Open Funder Registry DOIs, RINGGOLD organisation identifiers and grant numbers. The method and prototype build on work completed previously in THOR around extensible identifier and relationship types (de Mello et al., 2016).

3.1.2 Scalability Requirements

Use of ORCID continues to grow, both in terms of researchers and API users. As we incorporate more identifier types, more links and more functionality, it is necessary to address the questions of scalability, resiliency and redundancy. Alongside this increased general usage, we have also experienced issues where 'rogue bots' have threatened to take the registry and API down by flooding the search API with requests. Another key requirement therefore is to ensure that we are able to separate out API services to minimise impact of such threats. This will also enable future developments that depend heavily on network resources and infrastructure, such as per-identifier webhooks (Webhooks Notifications, 2017).

3.1.3 Search Extensions

In addition to the active linking service, ORCID EU also worked to address a search use case: "as a researcher/institution/funder I would like to discover all the ORCID records that are linked to a particular ID so I can track and report outputs". This use case was met by extending and improving the existing search API to work with organisation identifiers, funder identifiers and grant identifiers, as well as the names of each.

3.1.4 API Improvements

In a similar way to earlier developments in the project that addressed inconsistencies in the ORCID API for external identifiers, work was undertaken to improve the consistency of affiliation representations to ease integration and interoperability with funders.

Grouping work and affiliation metadata records into single entities requires ORCID to ascertain whether two identifiers are equivalent. However, despite past work on identifier representations, this is often difficult to establish. A clear case can be illustrated with the following:

- http://doi.org/10.1/1a
- http://dx.doi.org/10.1/1a
- https://dx.doi.org/10.1/1a
- doi:10.1/1a
- 10.1/1a
- 10.1/1A
- etc.

Other identifier types offer different challenges. For example, ISBNs can be represented in three ways, all of which are valid:

- 1234567890123
- 123-456-789-012-3
- 123 456 789 012 3

Many systems produce values that may not be technically correct, such as “ISBN-10: 0470860782”, but which must also be considered and parsed correctly. Work is continuing on how ORCID deals with the equivalence problem.

3.1.5 User-Facing Improvements

In addition to institutional users, over 3.2 million researchers have registered for an ORCID identifier. Their interoperability needs must also be addressed.

A user interface requirement emerged from the extended identifier types that we developed as part of previous work in THOR (de Mello et al., 2016), and we have addressed this by adding autocomplete typeahead identifier selection with the most common identifiers being easily accessed first. This replaces the dropdown list implemented previously, which was becoming unwieldy due to its size.

3.2 EMBL-EBI

The requirements for EMBL-EBI implementation follow three directions:

1. To enable life science repositories to assign DOIs to datasets alongside other identifier systems, and align these systems with DOI-based services as far as possible
2. The development of PID-based cross-linking services that identify relationships between archived data, metadata and research articles
3. The implementation of prototype services to link works claimed retrospectively into ORCID records to both funding and affiliations

All data resources at the EBI repositories use custom identifier systems (accession numbers), most of which predate the use of DOIs for data − in some cases by decades − and are instantly recognisable to the researchers that use them. Given this history, and the heterogeneity, complexity and scale of life sciences data, the question of assigning DOIs to these data applies only in certain cases.

The data resources at the EBI consist of both (1) submission databases (which accept data from the experiments of researchers), and (2) added-value databases, which can be built on top of the submitted data to provide more user-friendly views. These two types of database are fundamentally different: while the data in submission databases are collected from the community and are relatively immutable, the records in added-value databases are derived either by human curator efforts or computational means, or a combination of the two. As a result they can be highly dynamic, representing the best possible understanding of a given biological entity on a given day.

Some types of submitted dataset are voluminous, with one experiment representing millions of data points. The application of DOIs thus only makes sense for submission databases at the level of experiment, rather than for specific data points. Given the research communities (and workflows) that use these accession numbers on a daily basis, the assignment of DOIs is typically as an additional identifier alongside the traditional one. The decision to do so lies with the managers, governance and community that use a particular database.

The second requirement is driven by implementation of unified cross-linking services. Resource cross-links at the EBI are currently collated in Europe PMC via a variety of mechanisms; these include citation of research articles in data records, text mining, and links provided by life science data resources that reside somewhere other than the EBI. All of these links are available via Europe PMC APIs, albeit in a manner that reflects the evolution of these links rather than as a coherent service devoted to these relationships. The EMBL-EBI collaborates on the Scholix project to deliver a more straightforward method to consume literature-data cross-links, and is reorganising the API to provide a module that summarises such cross-links, regardless of their origin.

To satisfy the third requirement for the implementation of prototype services that link works claimed retrospectively into ORCID records to funding and affiliations, EMBL-EBI is exploring with ORCID how to represent funding and affiliation information in the metadata of individual works. If this can be done, then it will be possible to send this information to ORCID on claiming, should the user wish to include it. A more ideal approach would be to develop a tool that retrospectively enables the owner of an ORCID record to annotate the individual works within that record with funding and affiliation information; yet the development of such a tool is currently beyond the resources allocated to this task.

3.3 PANGAEA

Abstracting from the specifics of individual systems such as data repositories, cross-linking identifier types at PANGAEA (and Earth and Environmental Sciences more generally) relies on a few key requirements.

First, cross-linking relies on a certain maturity level of PID systems and workflows where, at a minimum, PIDs are registered and resolvable. If lacking this requirement, integrating and cross-linking PIDs in PANGAEA will result in a suboptimal user experience that suffers from limited discoverability of related contextual information. As cross-linking efforts at PANGAEA demonstrate, not all PID systems are equally mature. For some, such as the International Geo Sample Number (IGSNs), PID resolution is not guaranteed (for example, the PID may have been generated but not registered); cross-linking is thus challenging. Even more challenging is cross-linking entities for which a “standard” identifier has not emerged, such as organisations or projects.

As information about cross-links between identifier types is more valuable when shared among systems and infrastructure, cross-linking performed at data centres greatly relies on infrastructure support for sharing link information between entities, including entities other than data, articles, and people. The RDA/WDS Scholarly Link Exchange Working Group is a collaborative effort (also involving PANGAEA) that tries to improve the state of the art in sharing link information for data and articles. Infrastructures are already looking beyond these entities. For instance, ORCID is concerned with the links between people and funding/organisations, and DataCite focuses on linking funding and affiliations to works. Such links are certainly relevant to PANGAEA, and Earth and Environmental Science data centres more generally, but are arguably not of primary importance to Earth Science research workflows. Here, information about links between other entities (for example, data/people and instruments and their models, data/people and samples, or data/people and software and models) is crucial. Cross-linking at PANGAEA, and disciplinary data centres more generally, rely on infrastructures that support use cases other than bibliometrics.

Third, the quality of data and metadata archived at PANGAEA, and the quality and consistency of cross-links between identifiers in particular, strongly relies on informed data curation processes. Informing curators regarding the existence of identifiers, instructing on best practices for how to cross-link identifiers is thus an additional and important requirement for the PANGAEA data centre.

3.4 CERN

CERN's efforts in cross-linking persistent identifiers centre around CERN Analysis Preservation, an internal tool for linking components of a HEP analysis together in a central place, to facilitate access and reuse, and to assist with internal review procedures. Cross-linking of identifiers is key to the primary functionality of this tool, adding value by enabling the convenient inclusion of analysis components from internal resources within the experimental collaborations. Specifically, a primary requirement for this work is an autofill feature that leverages connections with collaboration-specific resources to populate the analysis metadata submission forms that a user encounters in CERN Analysis Preservation.

Additionally, CERN Analysis Preservation must have the flexibility to accommodate collaboration-specific resources that vary in the scope and types of data they store, and that store these data in a variety of formats. Collaborations operate independently and develop their own means of cataloguing and tracking analysis components across multiple collaboration-internal systems, varied by material type. For example, there may be one collaboration-specific resource for datasets, another for tracking software, and another to track internal review functions. These resources have their own internal identifiers. These internal identifiers can be leveraged to allow for the automatic harvesting of metadata into CERN Analysis Preservation from the relevant internal collaboration-specific resources that will ease the user data entry burden via the autofill feature mentioned previously.

CERN Analysis Preservation therefore serves to unite all of the system-specific internal identifiers that point to individual pieces of an analysis into one cohesive conceptual whole. In this way, CERN Analysis Preservation is in line with the strategy taken across all of the current CERN information products, serving as a unifying hub for multiple units, formats, and versions of related content, but on a larger scale and at a place farther upstream in the overall data workflow.

Despite being a single internal tool meant for use by all, the particular experimental conditions at CERN require that there be strict separation and access control between the various experimental collaboration groups. Collaborations, and even individual work groups within those collaborations, must have control over access to their analyses. This is a requirement that presents a special challenge, particularly on the social/political side of encouraging collaborations to link with the tool. This challenge is described further in section 5.4.

3.5 DataCite

A number of requirements for DataCite and DataCite Services have emerged while addressing this task, which overall consist of implementing prototype services for linking funding and affiliations to works claimed retrospectively.

Before outlining these requirements, we remind the reader of the concepts underlying DataCite's schema. The DataCite schema has at its centre the concept of a 'resource'. A resource includes academic works, such as a publication, a code repository, or a research dataset. The schema models the relation-ships of a resource with other objects. For example, the schema captures the relationships between a research dataset and all the publications that refer to it. Therefore, the focus of DataCite services is around resources. These internal services try to populate, display and enable external services that seek to interact with this model.

The first requirement that DataCite faces is related to the maturity of external services that could be used for the integration and cross-linking of both affiliations and funding. There are a number of services that could be used to provide integration and cross-linking of affiliations, but the lack of maturity of organisational ID services leads us to think that we should wait for the work on organisational identifiers to develop further. In the case of services that could be used to provide integration and cross-linking of funding, the Open Funder Registry managed by Crossref and using DOIs as a persistent identifier provides a mature persistent identifier service. Therefore, we address this requirement by scoping the implementation of the prototype to the integration and cross-linking of funder IDs from the Open Funder Registry.

A second requirement stems from the need to address the following use case: "as a researcher/Institution/Funder, I would like to discover all the works that are linked to a specific funder so I can track and report their outputs". In practice, this translates into the ability to search for works by the funder associated to those works. We address this requirement by laying the groundwork in the DataCite API for our services to cope with displaying the captured funding information and allowing authors to provide funder information for their ORCID record.

3.6 The British Library

ORCID and ISNI are separate, independent organisations that provide identifiers for individuals. They use the same identifier format, a 16-digit number string, although they use different number spaces meaning that the same 16-digits cannot be both an ISNI and an ORCID. While ISNI and ORCID have different missions and employ different processes for assigning identifiers and linking people with their works and affiliated institutions, points of overlap between them justify supporting interoperability.

ORCID and ISNI agreed on the need for interoperation, signing a Memorandum of Understanding in 2013. One outcome of that agreement was the development of a prototype service for linking and sharing public data between the two systems, as part of the EC funded ORCID and DataCite Interoperability Network (ODIN) project. The ODIN prototype enabled linking to ISNIs from within ORCID[2]. The ODIN prototype was not designed for long-term use, and was suffering from unclear ownership (hosting and maintenance), an aging codebase, deprecated libraries and incompatibility with the most current version of the APIs it uses. In THOR, we developed an improved and updated version of this tool. We upgraded the tool to the newest version of the ORCID API and fully integrated it into the ISNI system, where it will be managed as part of regular ISNI maintenance. The new THOR tool now also makes ORCID iDs available within ISNI services and APIs. It enables users of ORCID and ISNI to locate their corresponding record in the other system and link them.

The upgraded tool provides a basis for expanding links to research works and datasets. This will support more effective and streamlined claiming workflows. For example, based on ORCID-ISNI linking, upcoming developments will resolve metadata for ISNI-associated works, such as listed ISBNs, and import the associated metadata to ORCID. There is a community appetite for linking services from researchers to repositories. A higher profile and better integration that leads to a more connected scholarly e-infrastructure with bidirectional links for individuals, benefits both ORCID and ISNI.

3.6.1 User Stories and Requirements

We implemented user stories for the following actors:

- ORCID User - a logged-in ORCID user going through the linking process
- ISNI User - a general public access user on the ISNI site
- Developer - Someone writing software that uses ISNI and ORCID metadata

The following user stories are supported:

- R1. As an ORCID User I would like to be able to initiate the linking process from within ORCID so that I can have a more complete scholarly record.
- R2. As an ORCID User I would like the user interface to present me with ISNI record results from all my name variations after initiating the linking process from ORCID so that I do not have to manually re-enter search terms and search multiple times.
- R3. As an ISNI User I would like to see ORCID iDs displayed within ISNI records on the ISNI website if a link is known so that I can navigate to the associated ORCID record.
- R4. As an ISNI User I would like to be able to search the ISNI registry for an ORCID iD so that I can discover which ISNI IDs are linked to it.
- R5. As an ISNI User I would like ISNI to authenticate ORCID users before adding their metadata to the ISNI registry so that the linked identifiers are authenticated and accurate.
- R6. As a Developer I would like to see ORCID iDs included in the metadata attached to an ISNI ID so that I can link the two within my own systems.
- R7. As a Developer I would like to be able to search ISNI for an ORCID iD using the ISNI search API so that I can write better software.

The following user stories were part of the original specification, but were removed from the tender to OCLC, as ISNI currently do not see themselves as a work metadata store:

[2] See https://figshare.com/authors/ODIN_Consortium/101695

- R8. As an ORCID User I would like to see a list of works associated with the ISNI IDs I have added to my ORCID record and be able to import them into my ORCID record so that I do not have to enter them manually
- R9. As an ORCID User I would like to see ISNI IDs and works that ISNI has already added to my ORCID record during the matching process be marked as such in the ISNI interface so I do not attempt to import them twice

Removing these user stories means that the new service does not support some of the functionality available in the earlier ISNI2ORCID, but it is appropriately aligned with the ISNI's current service model and strategic approach.

4 Results

4.1 ORCID

4.1.1 Prototype Funding-Work Link Tool

ORCID EU has developed a closed, proof-of-concept prototype web application which demonstrates the feasibility of using information linked to ORCID iDs to facilitate retrospective links between funding information and works. It reaches out beyond the ORCID record to find works in the DataCite, Crossref and Europe PMC databases. For the purposes of this proof-of-concept, we use the use case of a hypothetical Europe PubMed Central funder who seeks to match funding to works for reporting purposes *within their own systems*.

The prototype addresses two related user stories:

1. As a funder I would like PIs and researchers to report on which works were produced with the funding I gave them so I can see the results of the funding.
2. As a researcher I would like to fulfil my reporting obligations to my funders without having to fill in the same forms over and over again.

The prototype walks the user through a series of steps that result in links between a funding item (identified by an organisation ID and grant number) and a set of works (identified by a DOI or PubMed identifier).

1. The user authenticates themselves using the ORCID API
2. The user selects one of their funding sources, which the prototype discovers using the ORCID API. If none are found, it prompts them to add funding sources to their record either manually or using the UberWizard search and link tool.
3. The user selects works from a list. The prototype builds this list using the ORCID, DataCite, Crossref and Europe PMC APIs to extract articles and datasets that reference their ORCID ID.
4. The user links the selected works to the funding source.

The prototype displays the provenance of the information that it imports; in this case, the sources are Crossref, DataCite, Europe PMC, ORCID, or a combination. This information about where bi-lateral links do and do not exist is not readily available elsewhere and indicates the level of cross-pollination between these systems. Further research and discussion around the value of this information and ways it can be utilised to increase coverage are needed. This could fruitfully be achieved in conversation and partnership with other organisations, such as OpenAIRE, and commercial organisations, such as ResearchFish, who are also working on this issue.

It should be noted that this kind of retrospective – or *active* – claiming is seen as complementary to automatic updates and submitting authenticated ORCIDs during manuscript submission, and is not meant to be the sole means of linking these kinds of entities.

User journey

The user journey is a simple workflow. Below we show screenshots of the prototype in the order in which a user would see them:

1. User is asked to log in with their ORCID account

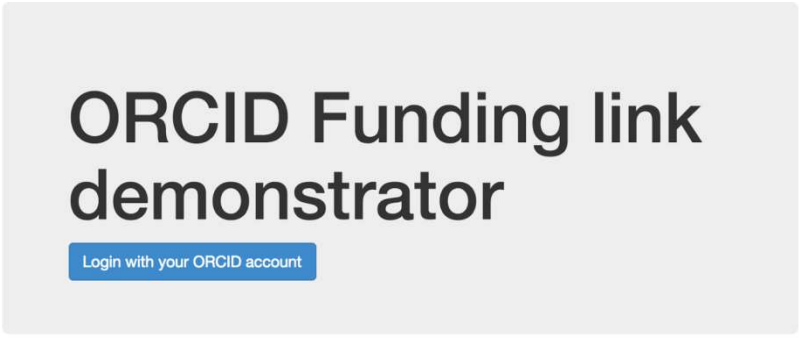


Figure 1: Login with ORCID account button

2. User grants permissions to the prototype

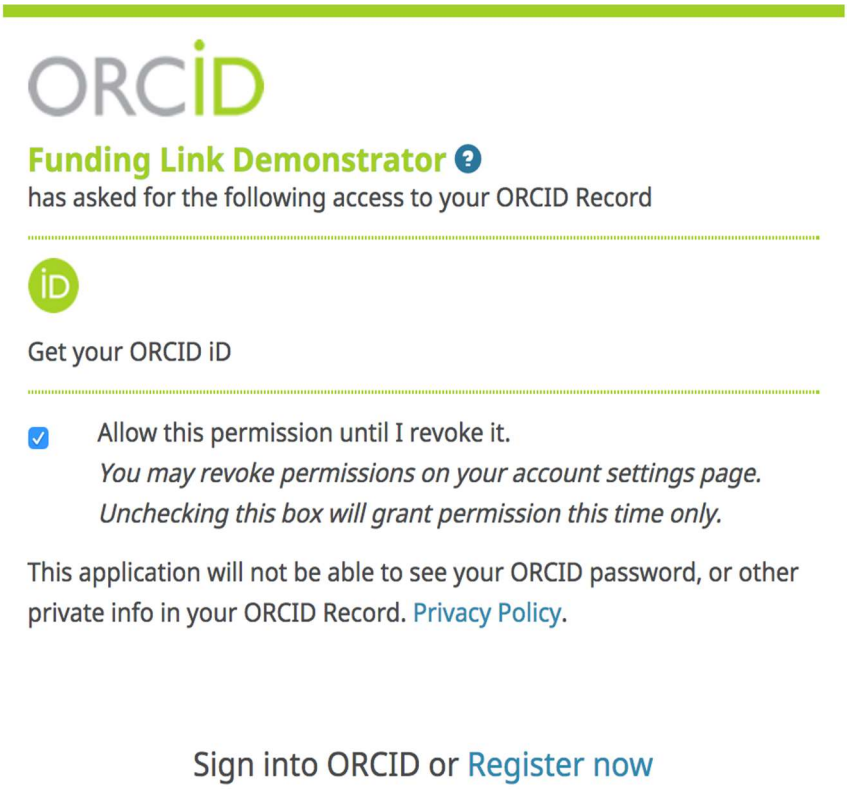


Figure 2: User authentication and authorisation on ORCID

3. User selects a funding source

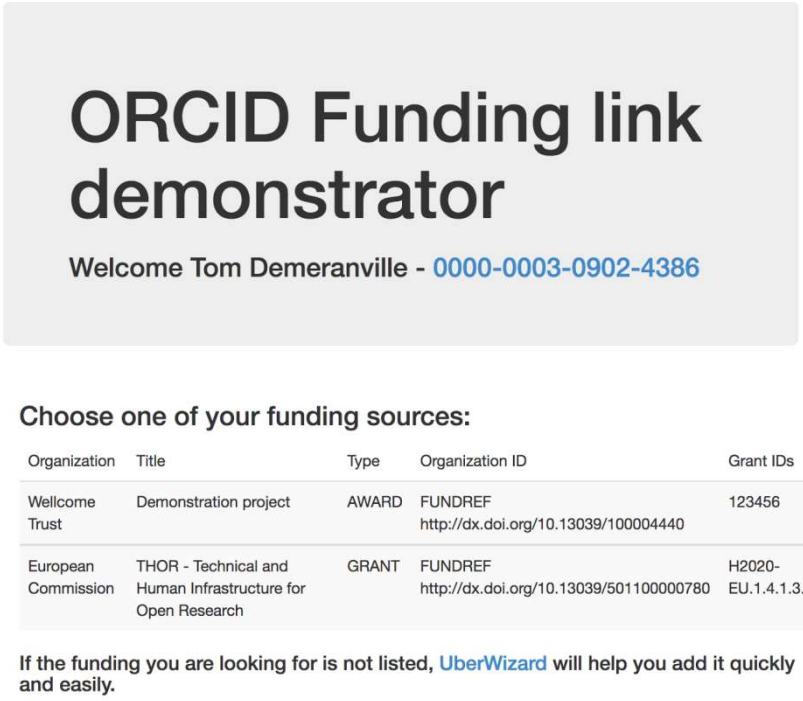


Figure 3: User selects a funding source

4. User selects works to link to the funding

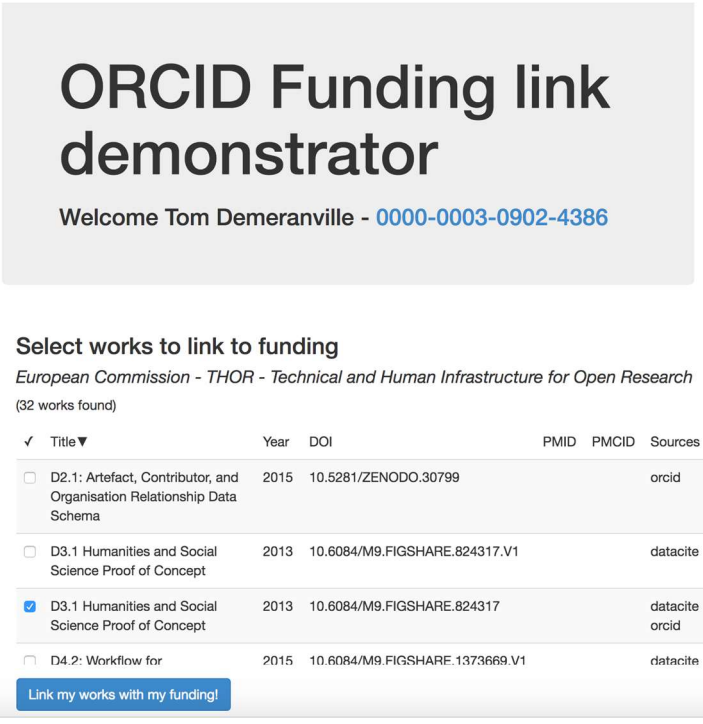


Figure 4: User selects works to link to the funding

Note that the prototype is implemented so that only a small change would be required to link works with other affiliations, such as Education and Employment. This is due to the changes made to the way Funding is expressed in the API, detailed below in the API improvements.

Persisting Funding-Work Links

There are several possibilities for the persistence of funding-work links, which could work together or separately. Options include ORCID, the funder identifier system (for example, Open Funder Data), the work identifier system (for example, DataCite/Crossref/PubMed), and the funders' own system. Another option is that a specialist system is created and designed to store funding-work links, in a similar manner to the way Scholix is designed to store data-article links. Further work is required to establish where this linking information should reside.

As described in the requirements section, ORCID does not currently support direct linking between two elements within an ORCID record. However, implementing a method to store links within ORCID, *as funding identifiers within work metadata*, is relatively simple. Funding identifiers such as DOIs, grant numbers and RINGGOLD could be attached to the works using the "part-of" relationship, which is semantically close to the intended meaning. This functionality is available now. The linking application would add a version of the works metadata with the funding identifier attached to the ORCID record using the ORCID API. Further work needs to be done to refine the relationship type, for example adding a more accurate "funded-by" type specifically for this use case.

Note that if a work already exists, then the user would have two versions of the work's metadata on their record. ORCID groups these works together with the full bundle of identifiers available via the UI and API. The source of the new work metadata would be the linking application.

4.1.2 Scalability

Before work could begin on adding further search indexes and improving the timeliness of search results, issues around scalability had to be addressed. Extensive analysis, design and development effort was spent ensuring that our server infrastructure would be able to support the increased load that populating and searching these indexes would generate.

ORCID search is powered by SOLR, which sits on a separate server. While this has the potential to be a bottleneck in the search process, this obstruction can be mitigated by scaling the service by simply assigning more computing resources to the SOLR server. The more difficult aspect to scale well is the method by which populating these indexes is achieved in real time. Previously, individual registry servers would poll the database periodically, looking for records that had changed in the last time period of, for example, one hour. It would then allocate resources to pull these records from the database and update the SOLR indexes record-by-record. The problems with this approach are that:

1. Updates are not performed in real time, meaning there is a disparity between the information in the registry and the ability to find it.
2. The main load on the infrastructure is on the same machines that are serving registry UI and API requests. This means that allocating resources to indexing takes it away from the regular registry functions.

3. Several registry machines running in parallel generate redundant database traffic.
4. Maintaining a complex code base that supports registry functions, search indexing and other "on-update" behaviour is difficult. It is harder to debug, improve and manage. Moreover, changing the way things are indexed requires a full release of our server stack.

We have significantly refactored the registry servers and established a message orientated middleware architecture based on the open source ActiveMQ. This enables our registry servers to "fire and forget" update notifications to a message queue, and to know that a machine elsewhere in our infrastructure is responsible for dealing with all actions related to update notifications, including search. A message listener machine has been put in place that takes these notifications and uses them to update the search index in real time.

The new message driven architecture is also the foundation we will build upon to provide services that will better enable large scale synchronisations across infrastructure providers. It is currently being used to populate a prototype 'real time data dump', which records all changes to the registry as a file store in the cloud. It will also be leveraged to improve our webhook functionality, and we are currently investigating the possibility of more granular webhook-delivered update notifications at the identifier (e.g. DOI) level as well as the existing per-ORCID level.

The version 2 search API was also modified to return lists of ORCID identifiers rather than full records to facilitate faster searching of larger datasets.

4.1.3 Search extensions

The scalability changes enabled us to add five new search indexes, which were created to facilitate looking up ORCID records by their funding and affiliations. The examples below return XML search results.

- RINGGOLD organisation identifier. Example: https://pub.orcid.org/v2.0/search?q=ringgold-org-id:385488
- Open Funder Registry identifier. Example: https://pub.orcid.org/v2.0/search?q=fundref-org-id:http%5C%3A%2F%2Fdx.doi.org%2F10.13039%2F501100000780
- Freeform grant numbers. Example: https://pub.orcid.org/v2.0/search?q=grant-numbers:H2020-EU.1.4.1.3.
- Affiliation names (Employment, Education, Funder). Example: https://pub.qa.orcid.org/v2.0/search?q=affiliation-org-name:orcid
- Funding titles/project names (e.g. THOR). Example: https://pub.orcid.org/v2.0/search?q=funding-titles:THOR

4.1.4 API improvements

The ORCID API currently supports three affiliation types: Education, Employment and Funding. The Education and Employment portions of the API have always been based on a similar metadata structure, enabling them to be treated in the same way for certain use cases. However, this was not the case for Funding. The main difference was that in order to resolve the organisation providing the funding, additional API calls were required. This was resolved by moving information about the organisation that provided the funding into the funding summary record; this is the same location as where the Education and Employment affiliation types are stored.

Work also commenced on managing the way identifier equivalence is calculated. This started with the ability to mark identifier types as case-sensitive or insensitive. Further work was undertaken analysing the way in which ISBNs are formed and compared, preparing the way for a solution that works well across all identifier types.

In addition, the following API-related improvements were made:

- New identifier types were added to the API and search indexes:
  - kuid - KoreaMed Unique Identifier
  - lensid - LENS Patent identifiers
  - ciencia-iul - As used by the ISCTE-IUL CRIS system for national reporting in Italy
  - igsn - International Geo Sample Number. Uses the new igsn.org resolver
- DOI creation behaviour within the user interface and API was modified so that new identifiers default to using HTTPS for URIs, unless otherwise modified.
- Behaviour of the peer review API was modified so that organisations could more easily discover which peer reviews they have added in the past so that they could better track and maintain their contributions to ORCID records.

4.1.5 User-facing improvements

Development effort was assigned to improve the way users select external identifier types to associate with their works. Changes were made so that the list was generated from the valid identifier types made available by the work undertaken in THOR previously (de Mello et al., 2016), with the most popular presented first.

4.2 EMBL-EBI

The THOR Project is one of the drivers behind the adoption of both ORCID iDs and DOIs for data at the EMBL-EBI, building on pre-existing resources such as the ORCID integration in Europe PMC, and components developed earlier in the THOR project such as the EBI ORCID Hub. The aim is to improve connectivity, discoverability and reuse of persistent identifiers, notably ORCID iDs and DOIs, through their delivery in current open services.

The EMBL-EBI is a Crossref member and thus can assign DOIs to content, including datasets and training materials. We undertook an informal survey to explore the potential use of DOIs and ORCID IDs within EMBL-EBI data resources. As stated earlier, all databases at the EBI use accession numbers, with a few submission databases (repositories) also assigning DOIs as a complementary identifier at the level of experiment (see Table below). Given the complex workflows and dynamic nature of knowledge bases, the application of both ORCID iDs and DOIs to EBI data resources will apply only to submission databases, in which the data are more stable and there is more often a clear data provider (author). However, it is worth noting that even archival datasets undergo maintenance updates, which are occasionally triggered by data source updates, format modification or changes to the data/metadata. Table 1 summarises the current situation at the EMBL-EBI regarding the use of DOIs in submission databases.

Table 1: Use of DOIs in EMBL-EBI submission databases

	Currently using DOIs	Would like to use DOIs for collections or high-level experiments
ArrayExpress		✓
BioStudies		✓
CHEBI		✓
ChEMBL	✓	for database releases
EGA		✓
ENA		✓
EMDB		✓
EMPIAR	✓	for experiments
EVA		✓
IntAct		✓
Metabolights		✓
Metagenomics		✓
PDB	✓ Used by RCSB not PDBe	
ProteomeXchange	✓	for experiments

(1) ChEMBL (see Figure 5) is a database containing binding, functional and toxicity information for a large number of drug-like bioactive compounds. These data are manually abstracted from the primary published literature on a regular basis, then further curated and standardised to maximise their quality and utility across a wide range of chemical biology and drug-discovery research problems. Occasionally, datasets are also submitted by researchers. The ChEMBL data can be accessed via a web-interface, RDF distribution, data downloads and RESTful web-services. DOIs are assigned for the small number of submitted datasets and for database releases (for instance, a new "edition" of the database) that happen periodically.

(2) PDBe (see Figure 6) is the European resource for the collection, organisation and dissemination of data on biological macromolecular structures. In collaboration with the other Worldwide Protein Data Bank (wwPDB) and EMDataBank partners, PDB works to collate, maintain and provide access to the global repositories of macromolecular structure data, the Protein Data Bank (PDB) and Electron Microscopy Data Bank (EMDB).

The assignment of DOIs to PDB datasets is currently performed by the US wwPDB partner (RCSB PDB). While the wwPDB partners all maintain their own portals, which often add value to the PDB data in different ways, they all share a single instance of the PDB data itself, which is available by FTP, and it is to this FTP site that the DOI resolves. This approach does not resolve the DOI to a landing page, but given that wwPDB consists of four partners, the FTP site is currently the fairest and most accurate place to resolve to. Currently the RCSB portal links from landing pages to the data via the DOI; subject to some technical issues being resolved, the PDBe portal at the EBI will be amending its landing pages to include the DOI in the near future.

Figure 5: ChEMBL DOI for a submitted dataset

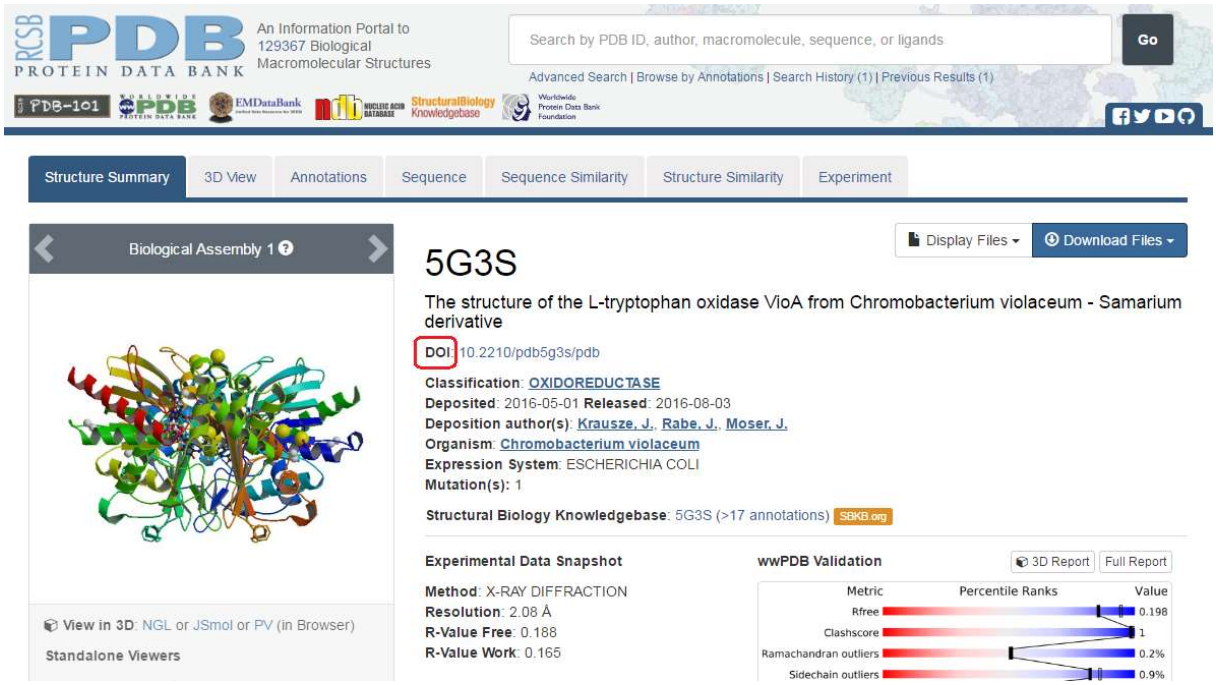


Figure 6: The RCSB PDB page lists the dataset DOI.

(3) EMPIAR (see Figure 7), the Electron Microscopy Public Image Archive, is a public resource for raw, 2D electron microscopy images. It complements the Electron Microscopy Data Bank (EMDB), where 3D images are stored. All EMPIAR entries have a DOI that links to the corresponding entry landing page. These DOIs have the format: https://doi.org/10.6019/EMPIAR-#####

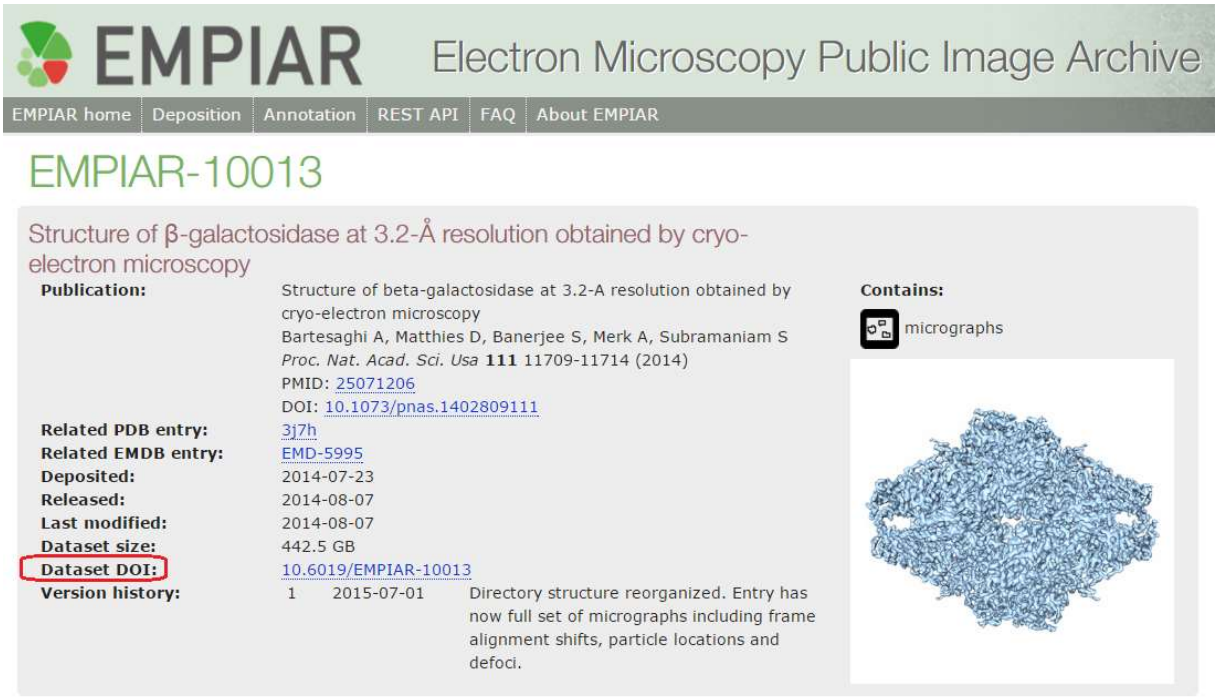


Figure 7: EMPIAR DOI association to dataset

4. PRIDE. The PRIDE PRoteomics IDEntifications database (PRIDE) is a core member of the ProteomeXchange (PX) consortium, which provides a single entry point for submitting mass spectrometry-based proteomics data to distributed public-domain repositories (mass spec datasets are very large). Datasets are submitted to PRIDE via ProteomeXchange and are handled by expert biocurators. Crossref DOIs are assigned at the experiment level, as requested by PX submitters.

4.2.1 ORCID Cross-Linking Across Multiple Data Resources

The EBI runs a cross-database search that is one of the most heavily used tools on the EBI website. While many life scientists bookmark the resources they use the most (therefore we are integrating ORCIDs into those specific resources), there is also merit in integrating ORCIDs into this cross-database search. EMBL-EBI is working on a pilot to enable ORCID integration into the cross-database search. This will take two forms: (1) enabling datasets already claimed to an ORCID to display the ORCID and also be discoverable by searching for that ORCID; and (2) allowing users to claim datasets to their ORCID across data sources, providing a “one stop shop” to claim datasets. This may be particularly useful for multiomic data providers, as the deposition of these data may occur across multiple resources.

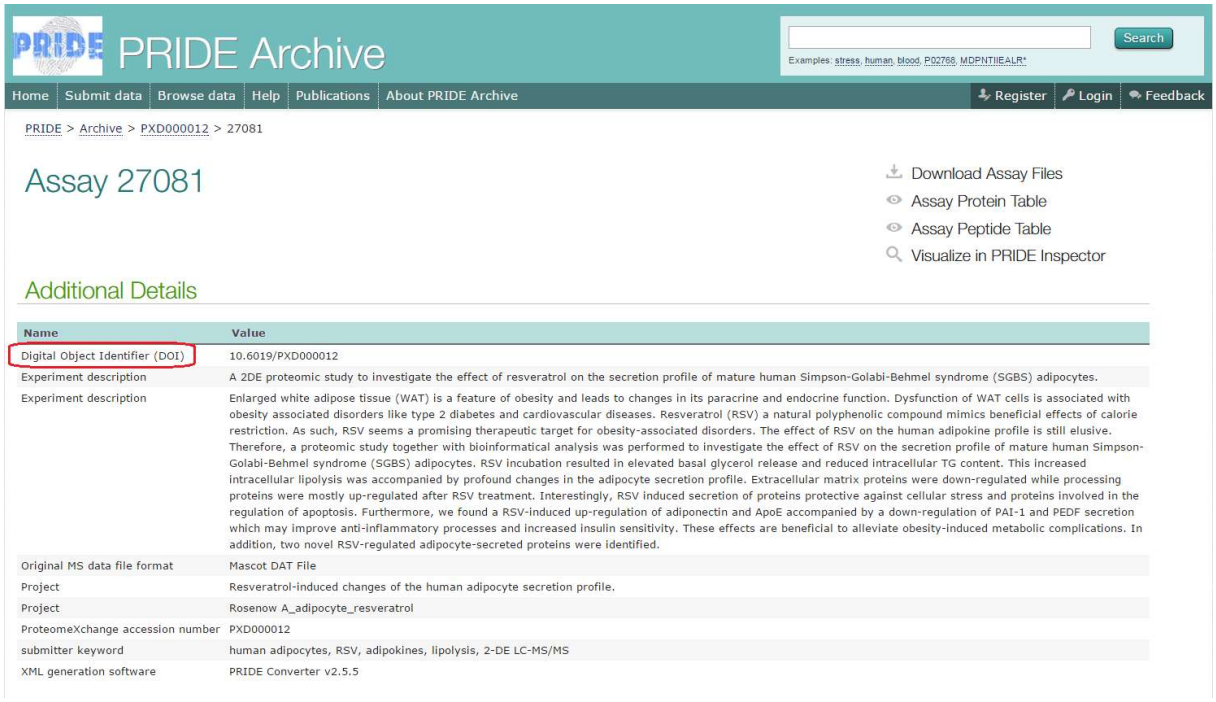


Figure 8: PRIDE landing page with DOI

1. Enabling ORCIDs in the EBI search interface: we will provide the EBI search with mappings of data accession number-ORCID pairs, plus the appropriate metadata, from the EBI ORCID Hub. This will allow an EBI-wide search for all EMBL-EBI datasets that have been claimed by that ORCID iD, and for all ORCID iDs that have claimed a particular dataset. As more datasets get claimed to ORCIDs, this will contribute to more widespread and deeper integration. The EBI Search will need to make a few modifications to the search results page to show the ORCID iDs for each entry.

The process for generating the data from EBI ORCID Hub for the EBI search is as follows (see also Figure 9):

- Step 1: the Automatic Generation Script responsible for generating the data for the search connects to the EBI ORCID Hub Web service at a specific URL.
- Steps 2 and 3: the EBI ORCID Hub starts the process of generating a mapping between the ORCID iDs and the datasets each person has claimed, from its own database records.
- Steps 4 and 5: the EBI ORCID Hub pushes the mapping information back to the Automatic Generation Script in JSON format.
- Step 6: the Automatic Generation Script exports the JSON file to an EBI Search XML format dump file, keeping the EBI Search up-to-date.

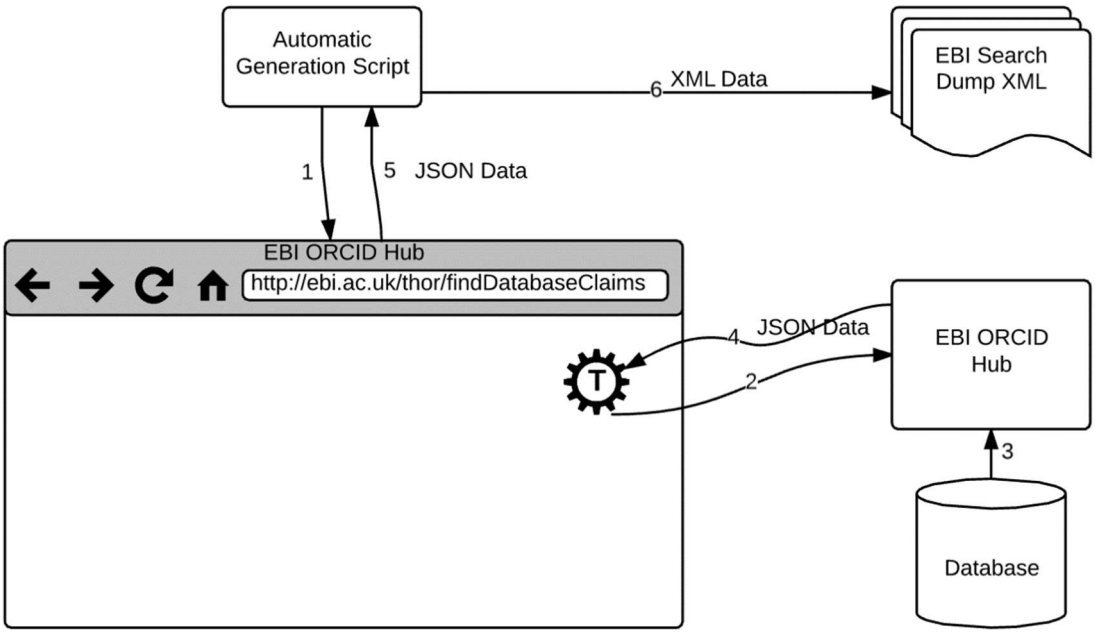


Figure 9: Generating the dump XML file to the EBI Search

2. Enabling batch claiming across databases, demonstrated by enabling ArrayExpress records to be claimed via a single 'Claim Selected Studies to ORCID' button on the EBI search result page. Conversely, users can also select which datasets they want to claim and use the 'Claim Selected Studies to ORCID' button to add all these datasets to their ORCID records, claiming several works at one time.

The steps required to claim works in batches involves the following (Figure 10):

- Step 1: the user clicks 'Claim Selected Studies to ORCID' button, and in turn the THOR Java-Script Client submits the data to the EBI ORCID Hub Batch Claiming Service.
- Steps 2 and 3: the EBI ORCID Hub redirects the user to ORCID for authentication and authorisation.
- Steps 4 and 5: ORCID directs the user back to the EBI ORCID Hub with an authorisation code.
- Steps 6 and 7: the EBI ORCID Hub exchanges the authorisation code for an access token.
- Step 8: the EBI ORCID Hub persists, in the database, the access token together with the corresponding ORCID iD.
- Steps 9 and 10: the EBI ORCID Hub uses the access token to push the claimed data to the user's ORCID record.
- Step 11: The EBI ORCID Hub Claiming Service informs EBI Search that the Dataset Claiming is successful.

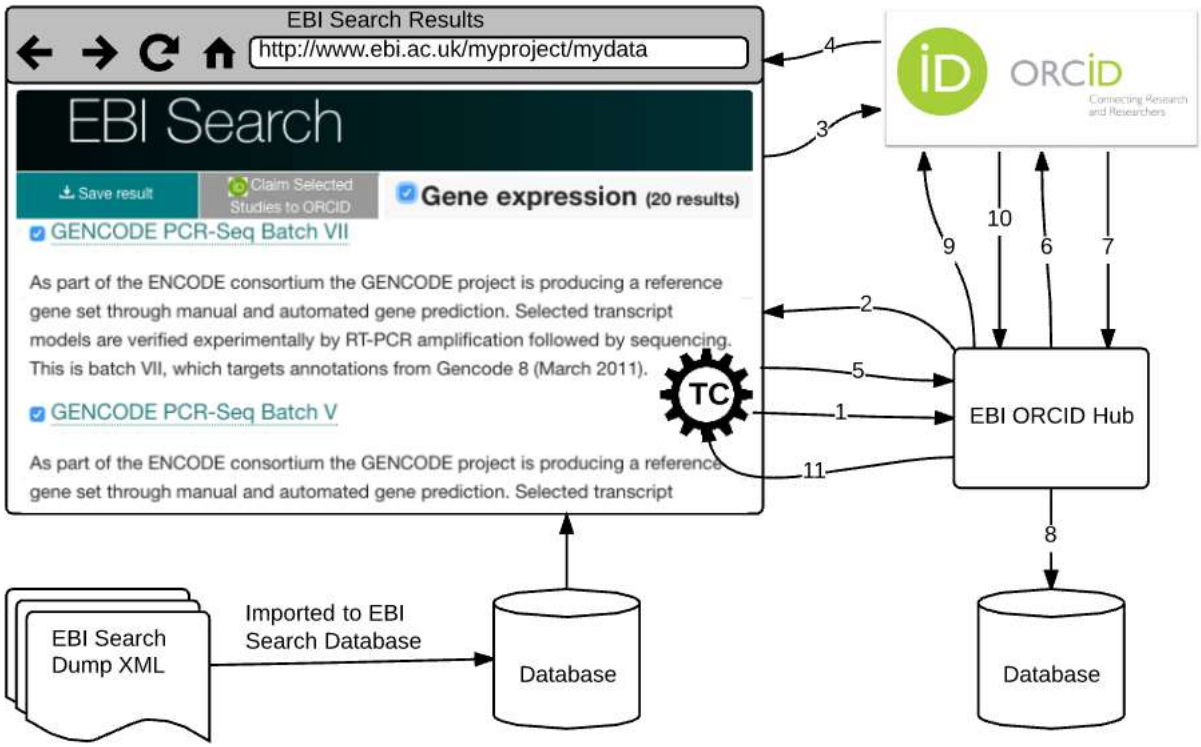


Figure 10: Batch dataset claiming with EBI ORCID Hub on EBI Search page

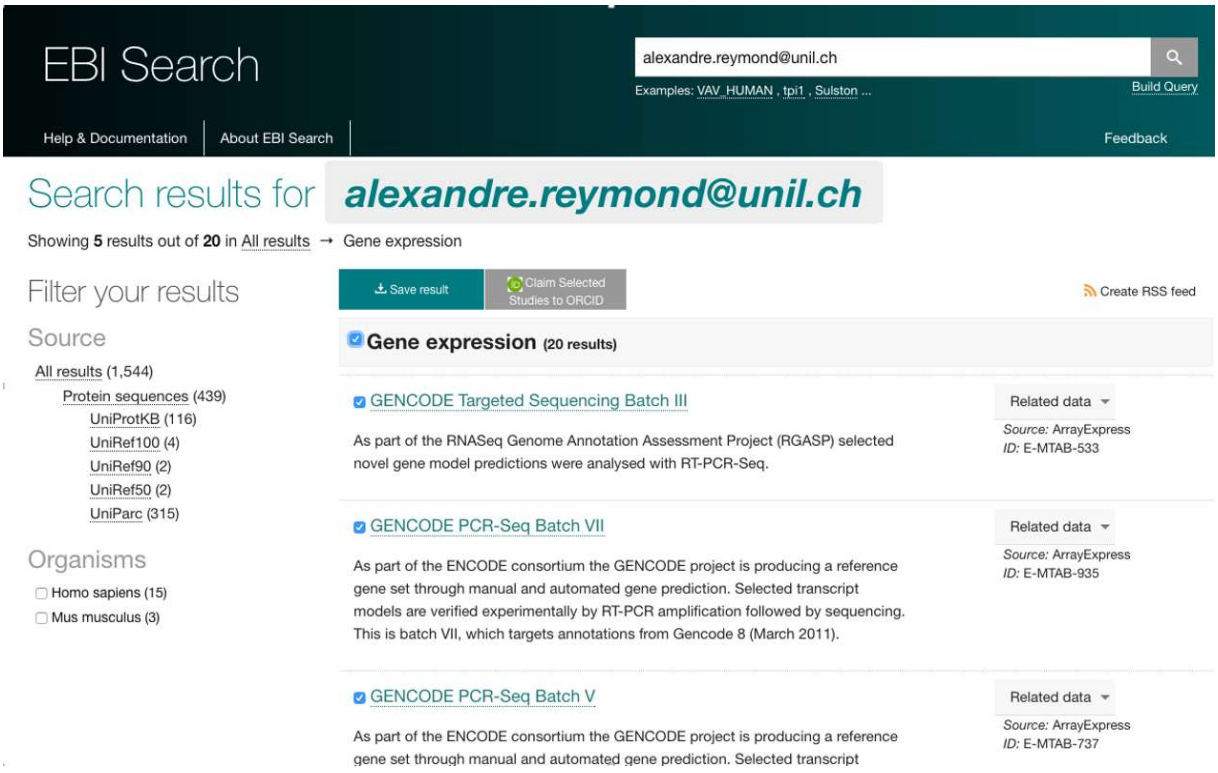


Figure 11: Batch claiming via 'Claim Selected Studies to ORCID' button on the EBI Search result page

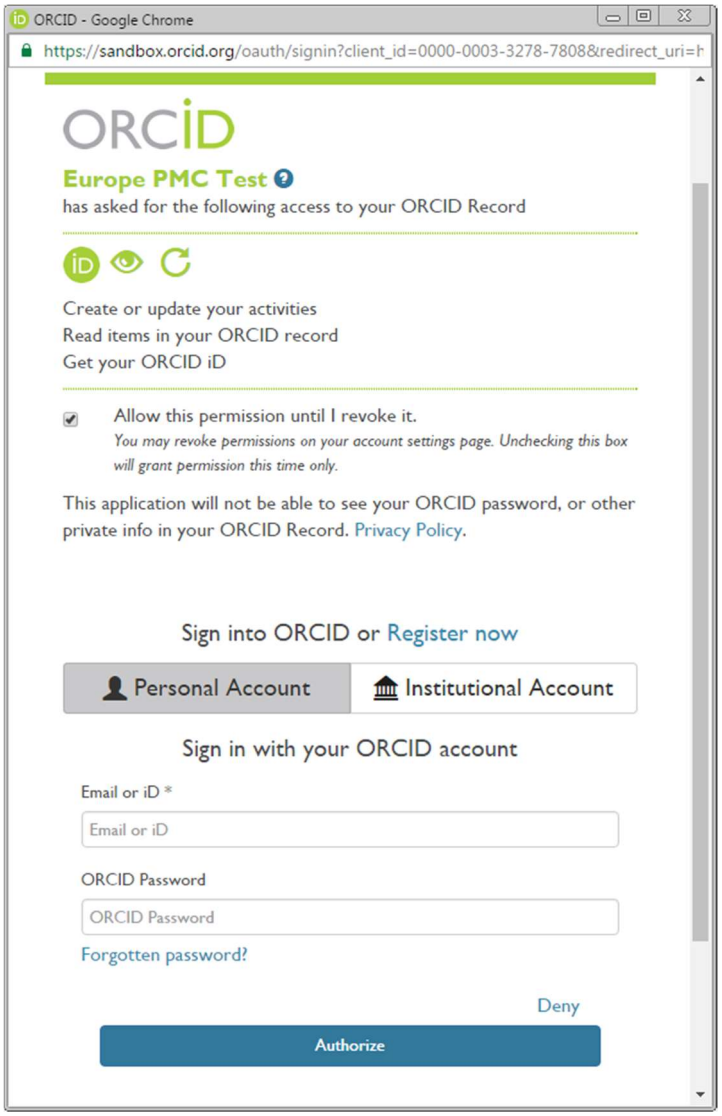


Figure 12: 'Claim Selected Studies to ORCID' button redirects the user to ORCID for authentication

4.2.2 An API Specifically for Life Sciences Literature-Data Cross-References

The Europe PMC APIs are used to populate the Europe PMC website and also directly, by thousands of users every month. These APIs are modular. They are organised into eleven modules, each serving different purposes:

1. Search - search, ranking and metadata retrieval of more than 30 million documents
2. Profile - summary of the different sources and types of documents for a given query
3. Citations - list of publications citing the input publication
4. References - list of publications that get cited by the input publication
5. databaseLinks - list of database records that cite the input publication
6. labsLinks - provides links to third party content that enriches a given publication
7. textMinedTerms - terms and accessions mined from open access publications
8. fullTextXML - serves the full XML of full text Open Access publications
9. bookXML - serves the full XML of Open Access publication of the Europe PMC bookshelf
10. supplementaryFiles - retrieve supplementary files in zip format
11. Fields - list of indexed search fields

Of these eleven modules, links from publications to data can occur within three, reflecting the method by which those links were obtained over time:

- Database links (databaseLinks): these are derived from the database's records of EBI resources when they cite a publication, which occurs when the data record was first published with the publication, or when the data record has been curated from a paper.
- Text-mined terms (textMinedTerms): this contains terms such as Gene Ontology terms, genes/proteins, diseases and species as well as accession numbers from over 20 databases identified through the Europe PMC text-mining pipeline. The accession numbers are identified first by pattern matching, supported by heuristics, and then resolved to specific data records, resulting in highly precise matching.
- External links (labsLinks): this module presents a variety of related content that ranges from article level metrics, lay summaries and press releases to data records in resources such as BioStudies or Dryad.

Since all three of these modules pertain to literature−data cross-links, they have been consolidated into a single data module.

This new data module provides resolvable data−literature links in a Scholix compatible format, a format drafted by the Scholarly link exchange RDA working group for the purpose of link exchange between natural link hubs such as CrossRef, DataCite and OpenAIRE.

The Scholix working group drafted guidelines and schemas to aid implementation of the Scholix-format.

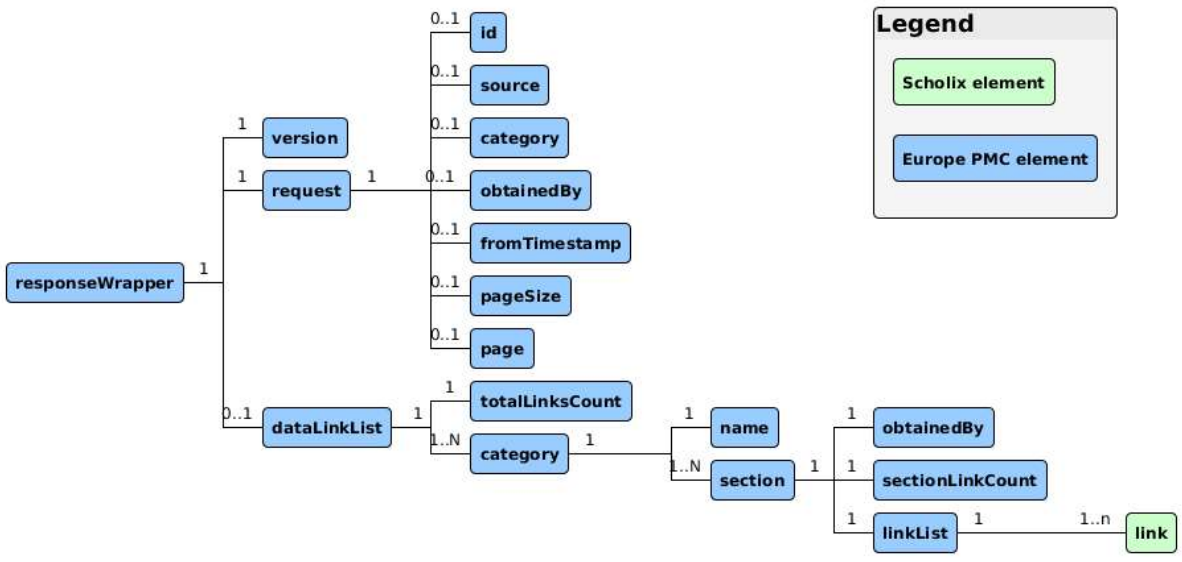


Figure 13: Europe PMC specific request metadata and categorisation

At the core of the Scholix hub infrastructure is the Scholix format. It provides a standardised container for data links that are agnostic of specific file formats like JSON or XML. This container also forms the centrepiece of the Europe PMC data module, yet it requires extension to match the richness of the prior Europe PMC webservice modules it will replace. For that purpose, the data module response groups the link information packages into higher level categories and sections, and adds extra elements.

The Scholix link element packages metadata relating to the relationships of literature and data entities. This includes directionality (for example, does the data record cite the publication or vice-versa?), a title of the source, and target entity and identifying metadata. The Scholix format was designed to be lightweight with very few mandatory fields. For the new data module, we considered it crucial to keep the size of link packages small. For some publications, there may be large numbers; transferring the packages may take so much time that it negatively impacts the user experience. The complete set of metadata can be reviewed in the Guidelines and the Schema provided by the Scholix working group[3].

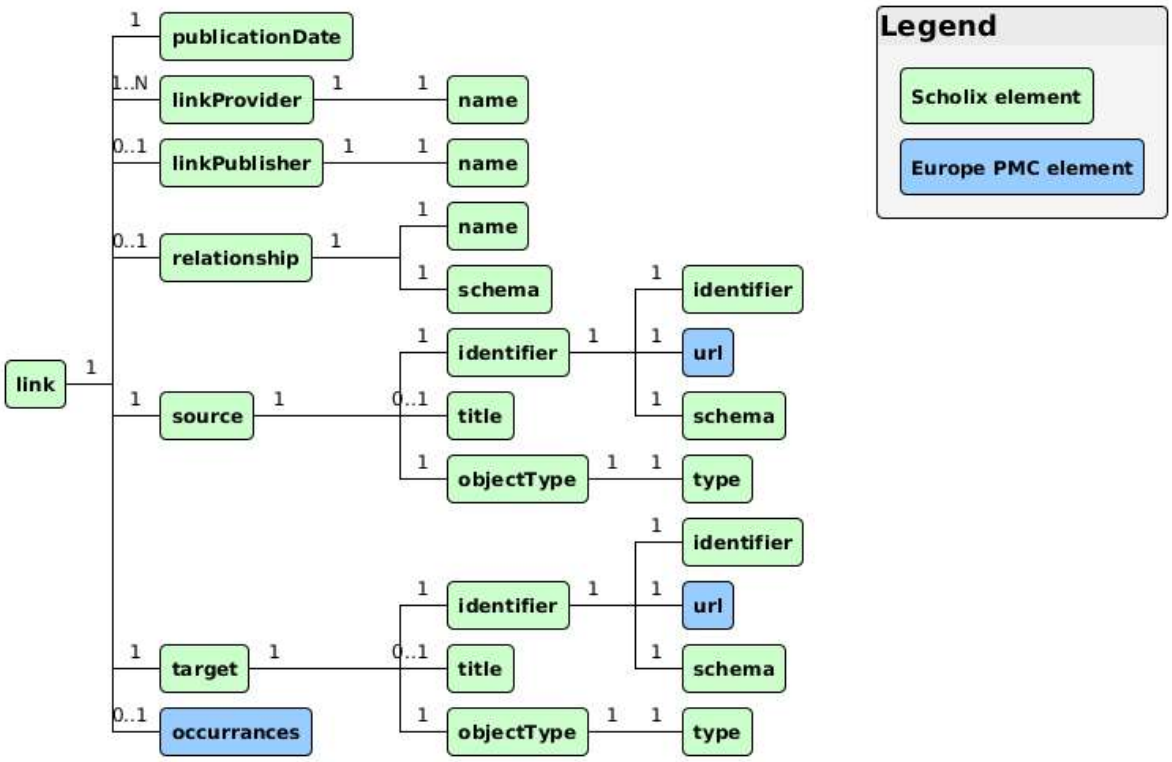


Figure 14: Scholix link format with Europe PMC extensions for the Europe PMC frontend

The Scholix metadata alone is, as mentioned, not sufficient to provide all necessary information to the Europe PMC frontend to allow linking directly. The Scholix links are meant to be atomic and do not allow two identifiers for a single object As a result, adding a URL as an identifier in addition the accession number would have forced us to duplicate the link information. To avoid this problem, we added a Europe PMC extension to allow URLs as an element in the Scholix identifier group. In the webservice response, this is implemented through the use of a Europe PMC and a Scholix namespace. Similarly, the occurrences

[3] http://www.scholix.org/guidelines

element is Europe PMC-specific. For text-mined terms and accessions, it provides the number of occurrences of the text-mined entity in the given publication.

This module was designed to serve content in a per-publication view, presenting data entities related to a specific publication, and also to provide data–literature links for third parties. To support the requirements of parties interested in literature–data links, we have also implemented an option to harvest these links incrementally and to be able to specify a category of interest. This is also in the interests of Europe PMC as a service maintainer because it reduces the necessary requests that harvesters have to make to keep their data up-to-date. To satisfy those requirements it is possible to query by:

- ID + source (data–literature links for a specific publication)
- Category (for example, PDB records)
- obtainedBy (for example, text mined terms, external links)
- fromTimestamp (limit the temporal scope of the request)

This allows efficient harvesting of data–literature links as it avoids bulk collectors having to request all possible links each time an update is required.

4.3 PANGAEA

Cross-linking identifier systems continue to be an active effort at PANGAEA. New identifier systems are developed by the community and, possibly, adopted by PANGAEA. Adoption necessarily lags behind the establishment of identifier systems. Furthermore, adoption is often incremental: PANGAEA may publish data that contain identifiers, and only later advance the system to recognise those identifiers as being of a distinct type.

The International Geo Sample Number (IGSN) is an interesting example. PANGAEA has long been publishing data and metadata that include IGSNs. A recent example can be seen in https://doi.org/10.1594/PANGAEA.857521. The data contain IGSNs, for example, CAI000001, as Sample ID parameter. Figure 15 visualises an extract of the dataset.

1 Event	2 Latitude	3 Longitude	4 Sample ID	5 Sample comment	6 Mg [# of ions]	7 SiO2 [%]	8 TiO2 [%]
AN160	19.1200	-99.7958	CAI000001	Toluca	53.7	64.8	0.67
AN179	19.1400	-99.8196	CAI000002	Toluca weathered	44.1	64.5	0.77
AN182	19.1400	-99.8273	CAI000003	Toluca	39.3	67.0	0.49

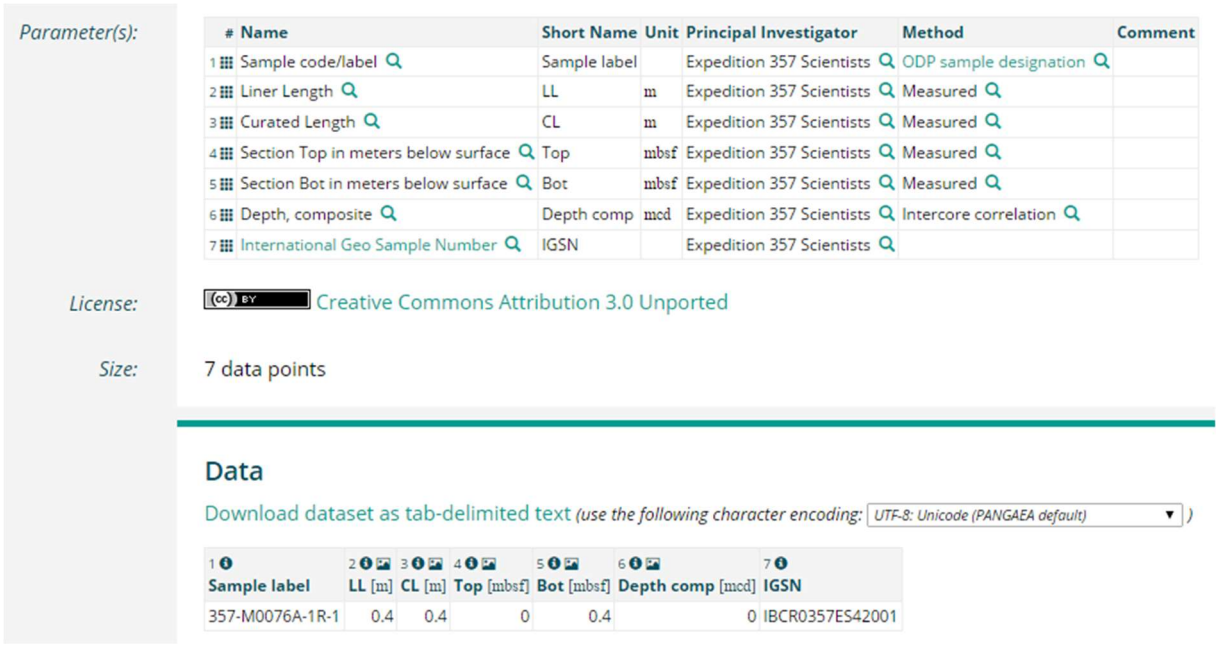
Figure 15: Use of International Geo Sample Number in PANGAEA.

Unfortunately, this representation has an important drawback. Even though IGSNs are Handles, and, assuming registration, they can be resolved to landing pages describing the sample, the representation here does not reflect this possibility by actively cross-linking the identifier. Knowledge of how to resolve

IGSNs is implicit and left to the user. A further difficulty is that the Sample ID parameter is not specific to IGSNs. Hence, Sample ID parameter values may be IGSNs, but often are values for some other kind of sample identifier system, including some devised by individual researchers.

Given these difficulties, PANGAEA has introduced a new parameter, IGSN, to distinguish such identifier types from others used to identify samples (Figure 16). It is thus possible to perform a PANGAEA search for datasets including the new parameter as follows:

parameter:"International Geo Sample Number"



Parameter(s):

#	Name	Short Name	Unit	Principal Investigator	Method	Comment
1	Sample code/label	Sample label		Expedition 357 Scientists	ODP sample designation	
2	Liner Length	LL	m	Expedition 357 Scientists	Measured	
3	Curated Length	CL	m	Expedition 357 Scientists	Measured	
4	Section Top in meters below surface	Top	mbsf	Expedition 357 Scientists	Measured	
5	Section Bot in meters below surface	Bot	mbsf	Expedition 357 Scientists	Measured	
6	Depth, composite	Depth comp	mcd	Expedition 357 Scientists	Intercore correlation	
7	International Geo Sample Number	IGSN		Expedition 357 Scientists		

License: Creative Commons Attribution 3.0 Unported

Size: 7 data points

Data

Download dataset as tab-delimited text *(use the following character encoding:* UTF-8: Unicode (PANGAEA default) *)*

1 Sample label	2 LL [m]	3 CL [m]	4 Top [mbsf]	5 Bot [mbsf]	6 Depth comp [mcd]	7 IGSN
357-M0076A-1R-1	0.4	0.4	0	0.4	0	IBCR0357ES42001

Figure 16: PANGAEA parameter for International Geo Sample Number (IGSN)

1 Event	2 Latitude	3 Longitude	4 IGSN	5 Sample comment
AN160	19.1200	-99.7958	hdl:10273/CAI000001	Toluca
AN179	19.1400	-99.8196	hdl:10273/CAI000002	Toluca weathered
AN182	19.1400	-99.8273	hdl:10273/CAI000003	Toluca
AN81	19.0930	-99.8006	hdl:10273/CAI000004	Toluca
M32	19.0550	-99.7264	hdl:10273/CAI000005	Toluca weathered
M34	19.0550	-99.7819	hdl:10273/CAI000006	Toluca
M69	19.0150	-99.7694	hdl:10273/CAI000007	Toluca weathered
RAM101	19.0980	-99.6616	hdl:10273/CAI000008	Toluca weathered
RAM215	19.1090	-99.7065	hdl:10273/CAI000009	Toluca weathered
RAM22	19.1180	-99.6602	hdl:10273/CAI00000A	Toluca
RAM452	19.1210	-99.6566	hdl:10273/CAI00000B	Toluca
RAM453	19.1060	-99.6567	hdl:10273/CAI00000C	Toluca weathered
RAM592	19.0850	-99.6725	hdl:10273/CAI00000D	Toluca

Figure 17: IGSN with active cross link to the Web location that describes the sample

This introduction also allows the active cross-linking of IGSNs in data (Figure 17). Unfortunately, supporting this feature for legacy data is not trivial. PANGAEA is currently developing a programmatic approach to identify IGSNs in data and accordingly update the parameter type in order to activate cross-linking. To programmatically identify IGSNs, PANGAEA is looking into utilising the list of Registered IGSN Namespaces.

A major challenge to cross-linking IGSNs is posed by the fact that not all published IGSNs are registered, and thus resolvable. This is because IGSNs are “minted in the field” and thus are used in the physical world but may not be registered in the digital world (for example, because registration is forgotten). PANGAEA is thus evaluating whether to activate cross-linking only if identified IGSNs are also resolvable.

At PANGAEA, cross-linking is a continued effort for many identifier types other than IGSN. The integration with Elsevier paved the way for data-article cross-linking. PANGAEA has been operating a Linking Hub lookup service that, given an article DOI, returns DOIs for published data that are supplementary to the published article (Figure 18). Given a PANGAEA data DOI, Elsevier can request a geospatial map widget to include in Elsevier article web views and directly link readers to supplementary data published by PANGAEA (Figure 19). Vice versa, PANGAEA informs users that published data are supplement to published articles (Figure 20).

Figure 18: PANGAEA Linking Hub lookup service for published supplementary data.

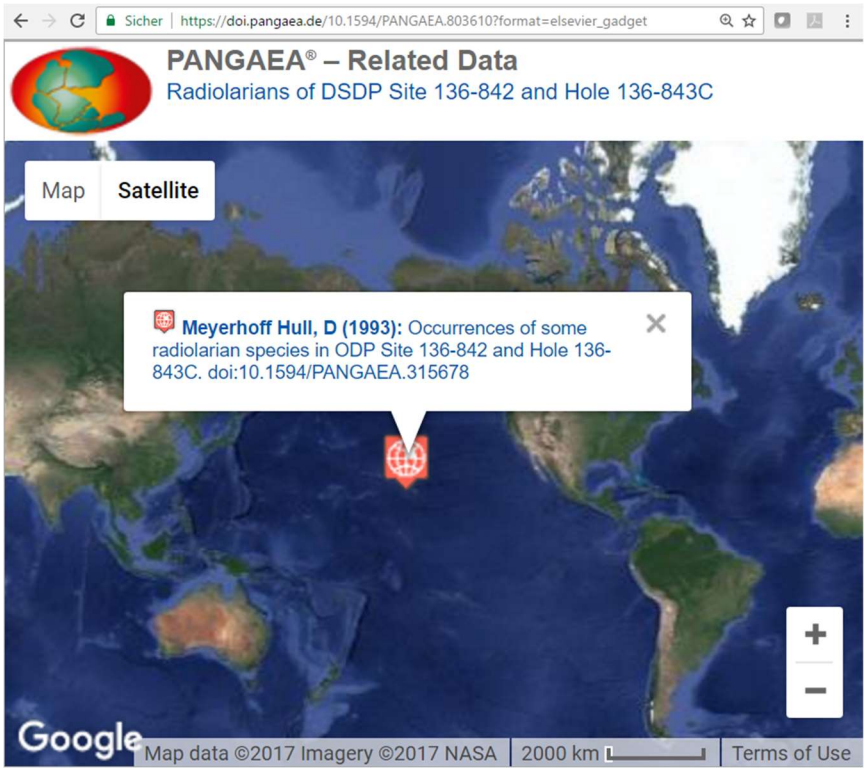


Figure 19: Map widget Elsevier includes in article views to cross-link supplementary data published by PANGAEA

Citation: **Meyerhoff Hull, Donna (1993):** Radiolarians of DSDP Site 136-842 and Hole 136-843C. doi:10.1594/PANGAEA.803610,

Supplement to: Meyerhoff Hull, D (1993): Quaternary, Eocene, and Cretaceous radiolarians from the Hawaiian Arch, northern equatorial Pacific Ocean. *In: Wilkens, RH; Firth, J; Bender, J; et al. (eds.), Proceedings of the Ocean Drilling Program, Scientific Results, College Station, TX (Ocean Drilling Program)*, **136**, 3-25, doi:10.2973/odp.proc.sr.136.201.1993

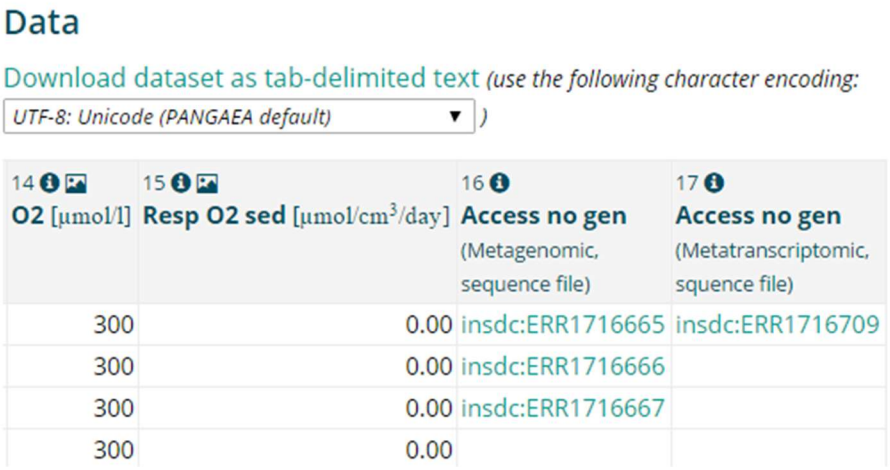
Figure 20: Cross-linking data (https://doi.org/10.1594/PANGAEA.803610) and article (https://doi.org/10.2973/odp.proc.sr.136.201.1993) in PANGAEA

In addition to data and articles, PANGAEA uses DOI cross-linking for data–data and collection–data. Such cross-linking is now gradually superseded by Scholix, which attempts to standardise the way that data centres, publishers, and PID infrastructures share information about the links between research articles and data. Such standardisation will simplify link information exchange, but it may no longer be possible to support some discipline-specific features, such as visualisation of geospatial map widgets.

PANGAEA cross-links entities other than data and articles, in particular accession numbers, campaigns, projects, institutions as well as terms of domain ontologies. However, cross-linking such entities remains more experimental and less standardised.

Regarding accession numbers, PANGAEA has started to cross-link such values found in datasets with descriptions provided by relevant databases. An example can be found in https://doi.org/10.1594/PANGAEA.867475, for which the data include the PANGAEA parameter “Accession number, genetics”.

As we can see in the data, values such as insdc:ERR1716665 are cross-linked with the corresponding description found at the European Nucleotide Archive. Figure 21 shows this functionality. Similar to IGSN and other identifier types, such cross-linking greatly increases the value to users as related information can be discovered easily and unambiguously.



Data

Download dataset as tab-delimited text *(use the following character encoding:*

UTF-8: Unicode (PANGAEA default))

14 O2 [µmol/l]	15 Resp O2 sed [µmol/cm³/day]	16 Access no gen (Metagenomic, sequence file)	17 Access no gen (Metatranscriptomic, squence file)
300	0.00	insdc:ERR1716665	insdc:ERR1716709
300	0.00	insdc:ERR1716666	
300	0.00	insdc:ERR1716667	
300	0.00		

Figure 21: Cross-linking accession numbers with the European Nucleotide Archive, an example for https://doi.org/10.1594/PANGAEA.867475; the value insdc:ERR1716665 cross-links with the resource https://www.ebi.ac.uk/ena/data/view/ERR1716665

Curators annotate metadata of datasets published by PANGAEA with information about campaigns, such as expeditions of research vessels. Campaigns are often identified by some kind of code and sometimes these codes can be cross-linked to web resources that further describe the related campaign. An example of such cross-linking can be seen in Figure 22.

Event(s):

PS93/050-5/6 (HG_IV) * *Latitude:* 79.065100 * *Longitude:* 4.181000 * *Date/Time:* 2015-07-26T00:00:00 * *Elevation:* -2465.5 m * *Location:* North Greenland Sea * *Campaign:* PS93.2 (ARK-XXIX/2.2) * *Basis:* Polarstern * *Device:* Multicorer with television (TVMUC) * *Comment:* Averaged Event from PS93/050-5 and PS93/050-6

Figure 22: Metadata annotation with information about the campaign associated to an event. The campaign code PS93.2 is cross-linked with a DOI-identified web resource that provides further information about the campaign. Here, cross-linking points users to a report on the expedition PS93.2 of the research vessel POLARSTERN to the Fram Strait in 2015 (https://doi.org/10.1594/PANGAEA.867475)

Similar to campaigns, curators also annotate metadata with information about projects and institutions related to datasets published by PANGAEA. Projects are typically identified by their CORDIS number, and cross-linked to the corresponding Web resource by URL. An example can be seen in Figure 23 for https://doi.org/10.1594/PANGAEA.842709. Institutions, such as MARUM − Center for Marine Environ-

mental Sciences, are sometimes part of metadata and may be cross-linked to web resources, such as the institution's website (see, for instance, https://doi.org/10.1594/PANGAEA.863306).

Figure 23: Cross-linking datasets published by PANGAEA and projects managed by CORDIS

Recently, PANGAEA also started to cross-link its terminology with concepts of external vocabulary. A prominent example is cross-linking of terms of the PANGAEA Feature Catalogue with the World Register of Marine Species. For instance, the term *Sotalia fluviatilis* – commonly known as the gray dolphin – can be cross-linked with the LSID urn:lsid:marinespecies.org:taxname:254982 – or, perhaps preferably, the URI http://identifiers.org/worms/254982. In the future, such cross-linking is expected to be extended from PANGAEA features to other terminology, including parameters and units.

These examples suggest that cross-linking practices vary depending on identifier type. While data–article or data–data cross-links are relatively mature, and fairly advanced infrastructure is emerging that supports the sharing of link information about these types, for other identified types more work is needed to harmonise practices, representation and communication. For instance, while most cross-linking is between persistent identifiers, in particular DOI but also IGSN, this is not the case for all cross-linking. As we have seen, projects and institutions typically cross-link to URL. This is partially due to lack of appropriate infrastructure (for example, for project identifiers), and partially due to curatorial practices. For institutions, it could be possible to cross-link using GRID or ISNI identifiers, rather than institutional website URLs. MARUM may thus be identified by grid.474422.3 or the ISNI 0000 0001 1013 246X. Different practices in cross-linking identifier types can be found also in representation. While some cross-linking is structured, for example, with a dedicated metadata attribute, other cross-linking occurs embedded in free text, such as for campaigns or projects. Finally, some identifier types suffer from the problem that not all instances of those types are formally registered and may thus not be resolved, even though the identifier is utilised and published. This issue is noted especially for IGSN.

4.4 CERN

CERN Analysis Preservation is an internal tool based on Invenio 3.0. Users from within the experimental collaborations that operate on the Large Hadron Collider (LHC) create and update records for analyses in CERN Analysis Preservation. In order for these analysis records to be useful for future collaboration members, a large amount of information must be collected regarding the features of the dataset used in the analysis, the processing steps carried out, and any relevant code. A particular challenge has been to balance the future usefulness of a high level of detail with immediate concerns of data entry fatigue. To help achieve this balance, we have sought connections with collaboration-specific resources to enable harvesting of relevant metadata based on internal identifiers.

CERN Analysis Preservation is currently connected to the internal resources of three of the four primary LHC experimental collaborations. We have set up service accounts specific to each collaboration, so that any queries for collaboration-specific information run across CERN Analysis Preservation, or future information services, is handled by a single, discrete account. This maintains some separation between collaboration information we collect, and any compromise of a single service account should have no bearing on any other collaboration.

User permissions are based on internal CERN user group permissions (e-groups). A user logging into CERN Analysis Preservation will be shown only those analyses affiliated with the e-groups to which they belong. These e-groups typically align with particular experimental collaborations and with specific working groups within those collaborations. E-group permissions are managed by the collaborations and working groups themselves, so they have complete responsibility for access control over their internal materials.

A user may choose to add an analysis for a particular experimental collaboration to which they belong. The user will be routed to the appropriate form built specifically for that collaboration based on an underlying collaboration-specific JSON schema that has been developed by the CERN Analysis Preservation team in consultation with representatives from the collaborations. The form is presented in sections as a way of making the volume of information entry easier to digest. Among the first fields at the top of the form, the user has the option to enter identifiers from those resources specific to the particular collaboration to which the analysis belongs. Information matching those identifiers is then fed into the form fields automatically, sparing the user the effort of manual entry.

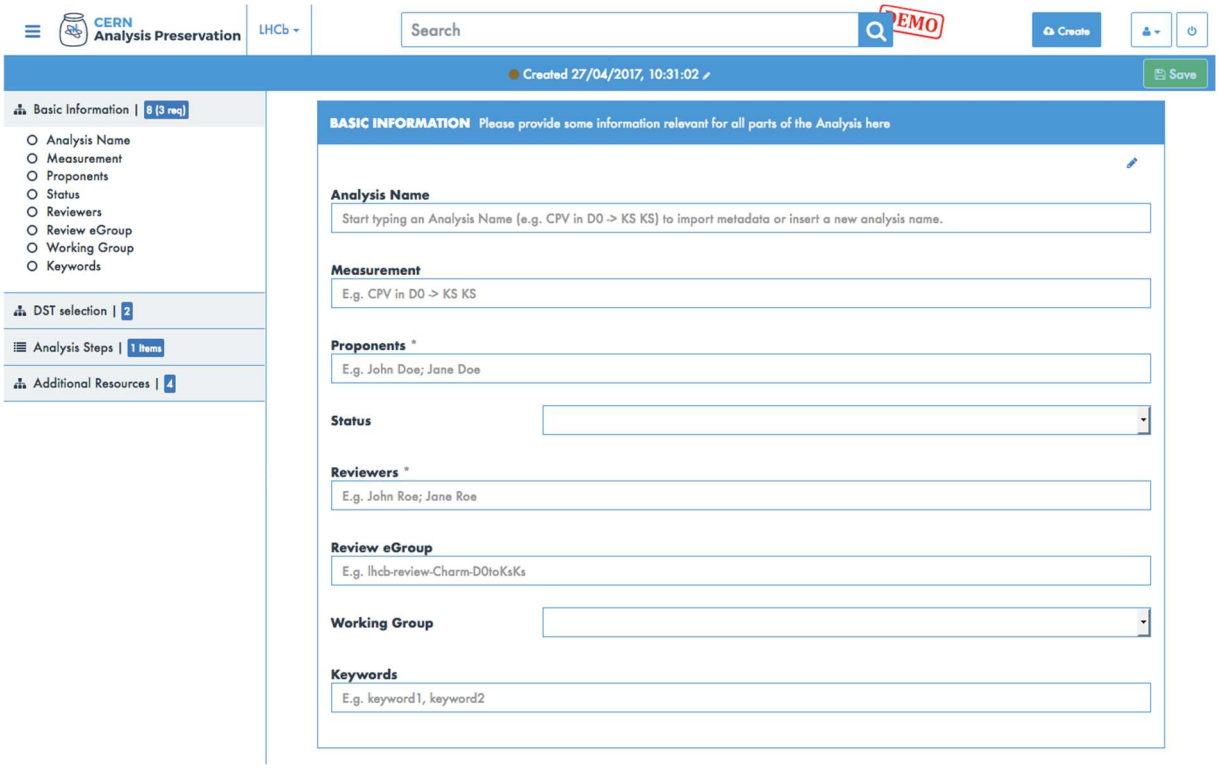


Figure 24: CERN Analysis Preservation submission form, simulating the view of a user in the LHCb collaboration

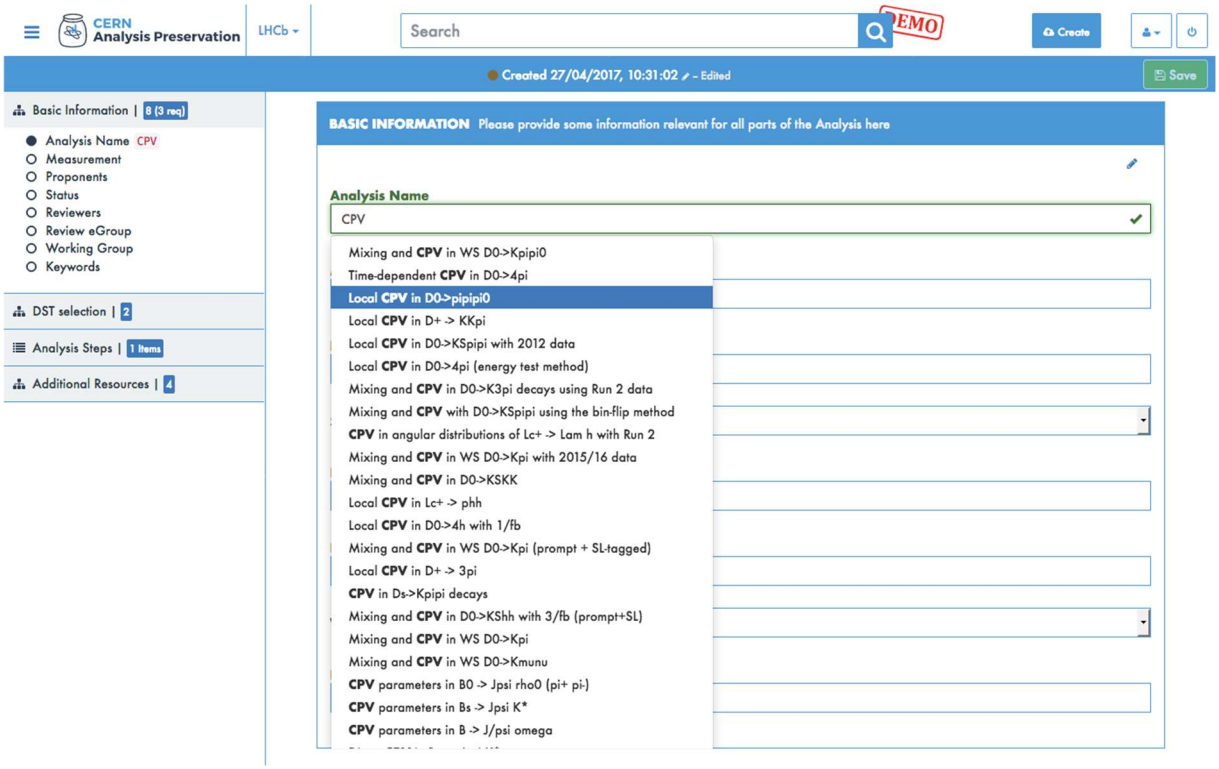


Figure 25: CERN Analysis Preservation autocomplete selections based on user input

Because we are connecting to multiple collaboration-specific resources, with varying levels of support for external applications, our approach to searching for matching information likewise varies. The ideal method is for us to perform on-the-fly queries of APIs specially prepared for us by the collaborations. This leaves the collaborations in control of the data shared with CERN Analysis Preservation, and it removes any of the issues that come along with syncing copies of stored data. In the absence of an API, a regular database query is a close alternative that provides much the same benefit. However, some of the collaboration-specific resources are stored in a format other than the expected relational database structure, such as a python shelf object. In these cases, we download the latest version of the files in these resources every night, pre-process them into an application-friendly format, and index them to enable CERN Analysis Preservation to search them via its regular Elasticsearch engine.

CERN Analysis Presentation is still in closed alpha, but initial feedback from collaboration representatives and test users has been positive. Users naturally appreciate being spared data entry effort, and re-using existing collaboration-specific information is seen as both a time saver and as a demonstration of respect toward established researcher practice.

4.5 DataCite

DataCite's effort was to implement the prototype service for linking funder information to works claimed retrospectively into ORCID records. Here we describe this effort in two parts. The first part deals with user-facing interactions that are involved in retrospectively adding funding information. The second part deals with the inner workings of DataCite services to facilitate the aforementioned interactions and machine interactions that other services can take advantage of.

In essence, the prototype focuses on the process that allows authors to retrospectively add funder information to their works. Most importantly, it focuses on facilitating this process while using PIDs (funder IDs from the Open Funder Registry) for integration and cross-linking. Similar to DataCite's previous implementation, which involved authors adding information to works retrospectively, in this prototype authors are able to perform such activity by visiting DataCite Search. Authors need to (1) authenticate in order to gain access to the prototype capabilities. As shown in Figure 26, a new button (2) at the bottom of each work enables authors to access these capabilities, and thus update funder information per individual work. Upon clicking the 'update funding' button, two sources of infor-mation are pulled to the front-end interface (see Figure 27). Those are (3) the names of the ~14,000 funders listed in the Open Funder Registry, and (4) funding information previously associated with the work in question. Authors are able to search through the list quickly for the name provided by Open Funder Registry using (5) a search box. They can select multiple sources of funding from the list.

The second part of the work implemented with this prototype involves all development work used behind the scenes to achieve funder ID cross-linking. There are a number of new endpoints in the DataCite APIs.

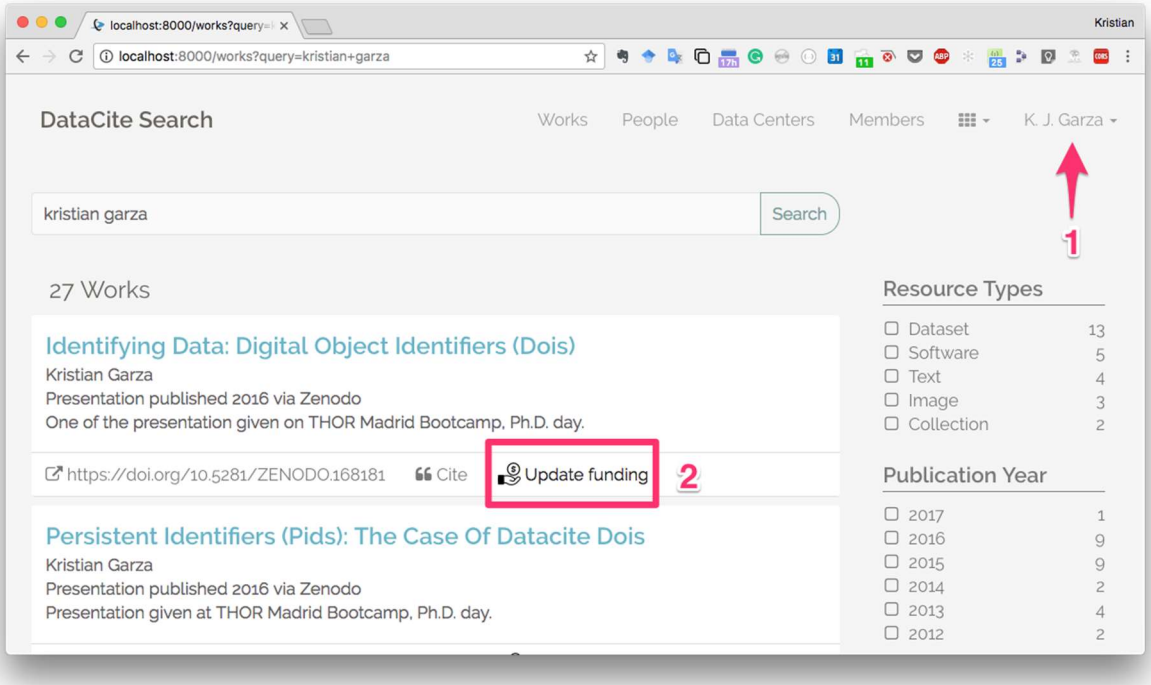


Figure 26: DataCite Search new update funding button

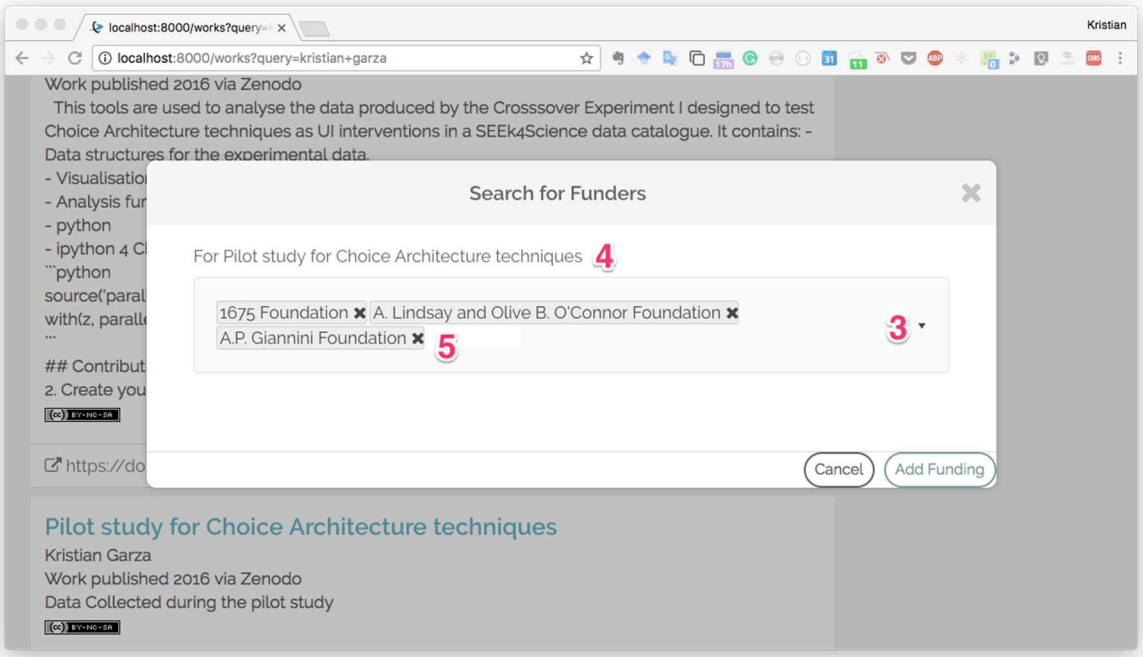
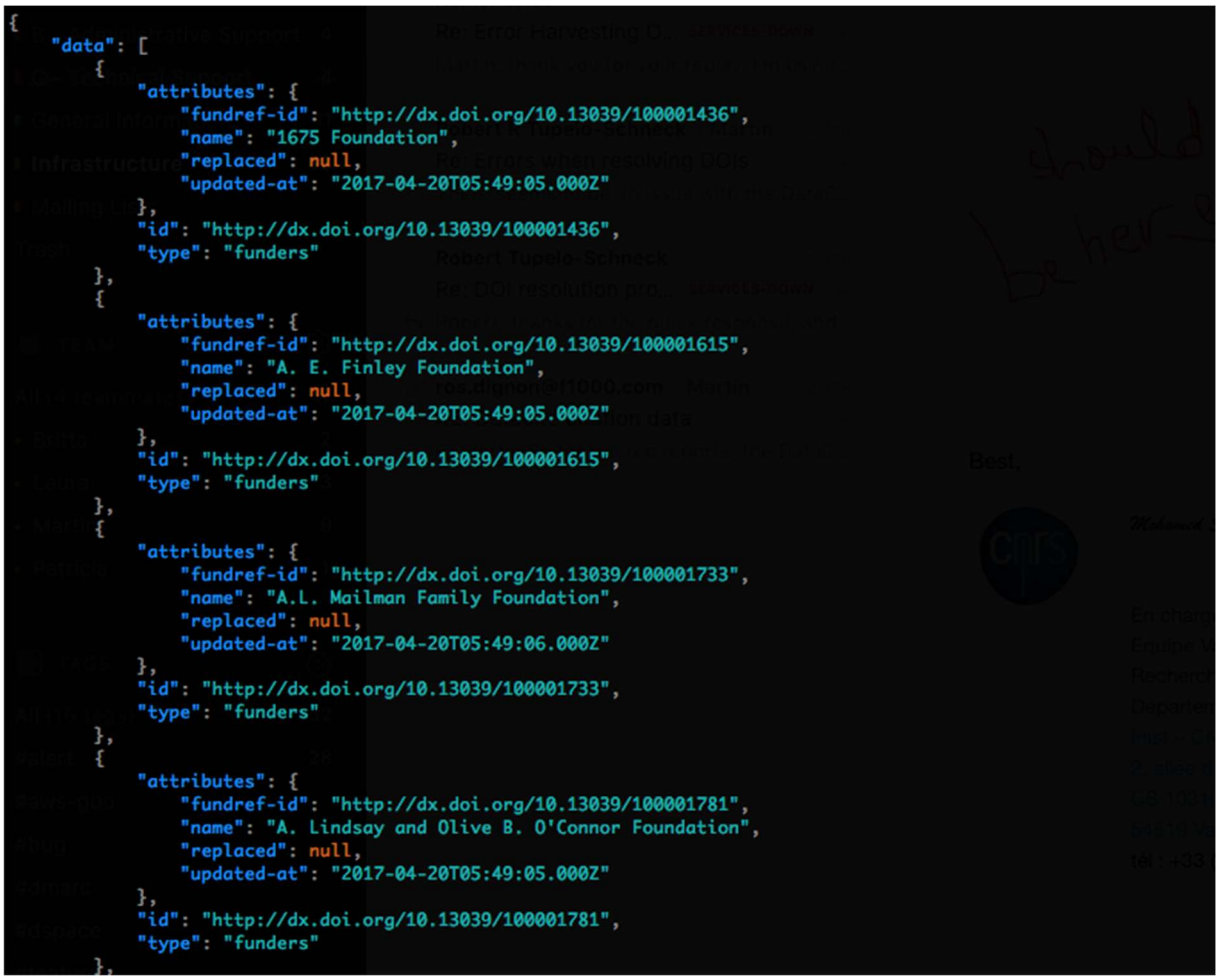


Figure 27: Authors can search the Open Funder Registry for funders to add to their works. Authors can add multiple funders to a work

4.5.2 Funders API

DataCite now provides a funders' endpoint that serves multiple use cases. For example, we use this endpoint to provide the DataCite Search front-end interface with funder names and funder PIDs. This API allows DataCite services to build deeper integrations going forward, for example a funder tab in our Search UI. The response to calling this API can be seen in Figure 28.



```
{
    "data": [
        {
            "attributes": {
                "fundref-id": "http://dx.doi.org/10.13039/100001436",
                "name": "1675 Foundation",
                "replaced": null,
                "updated-at": "2017-04-20T05:49:05.000Z"
            },
            "id": "http://dx.doi.org/10.13039/100001436",
            "type": "funders"
        },
        {
            "attributes": {
                "fundref-id": "http://dx.doi.org/10.13039/100001615",
                "name": "A. E. Finley Foundation",
                "replaced": null,
                "updated-at": "2017-04-20T05:49:05.000Z"
            },
            "id": "http://dx.doi.org/10.13039/100001615",
            "type": "funders"
        },
        {
            "attributes": {
                "fundref-id": "http://dx.doi.org/10.13039/100001733",
                "name": "A.L. Mailman Family Foundation",
                "replaced": null,
                "updated-at": "2017-04-20T05:49:06.000Z"
            },
            "id": "http://dx.doi.org/10.13039/100001733",
            "type": "funders"
        },
        {
            "attributes": {
                "fundref-id": "http://dx.doi.org/10.13039/100001781",
                "name": "A. Lindsay and Olive B. O'Connor Foundation",
                "replaced": null,
                "updated-at": "2017-04-20T05:49:05.000Z"
            },
            "id": "http://dx.doi.org/10.13039/100001781",
            "type": "funders"
        },
```


Figure 28: DataCite Funders API endpoint can be called at https://profiles.datacite.org/funders/

4.7 The British Library

The upgraded THOR ISNI–ORCID integration tool now satisfies the requirements that were specified in Section 3.6. This section describes the basic functionality that has been implemented for the stakeholder groups: ISNI users, ORCID users and developers.

4.7.1 For ISNI Users

The required integration functionality for ISNI users has been implemented as a tool, which is one of a family of integrations known as a 'Search and Link Wizard', which describes a type of ORCID integration that can be initiated from within the ORCID user interface. These wizards sit on third party infrastructure and follow similar patterns: obtain permission from the user to read/write to their ORCID record, decide what to add/update, perform updates, and redirect the user back to ORCID.

The basic workflow for adding an ORCID iD for an ISNI user has been implemented in the following way.

1. The user logs in to ORCID at https://orcid.org/signin.

   Logging into ORCID to access the ISNI–ORCID tool ensures that ORCID users are authenticated before adding their metadata to the ISNI registry, thereby satisfying requirement R5.

2. The user locates the 'Add works' button within the ORCID registry, and selects the 'Search & link' functionality.
3. The user selects the 'ISNI-ORCID' service from the list of wizards.
4. The user is redirected to the ISNI-ORCID service.
5. The user completes the ORCID OAuth flow where the service asks for permission to update their record. This authenticates the user and gains update permissions.

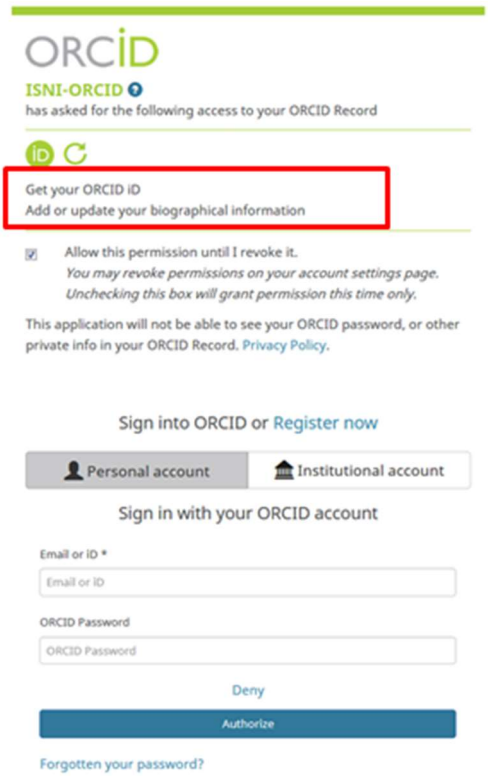


Figure 29: ORCID OAuth flow where the service asks for permission to update their record

6. The user is presented with a list of ISNI name search results, using their ORCID name(s) as the initial search.

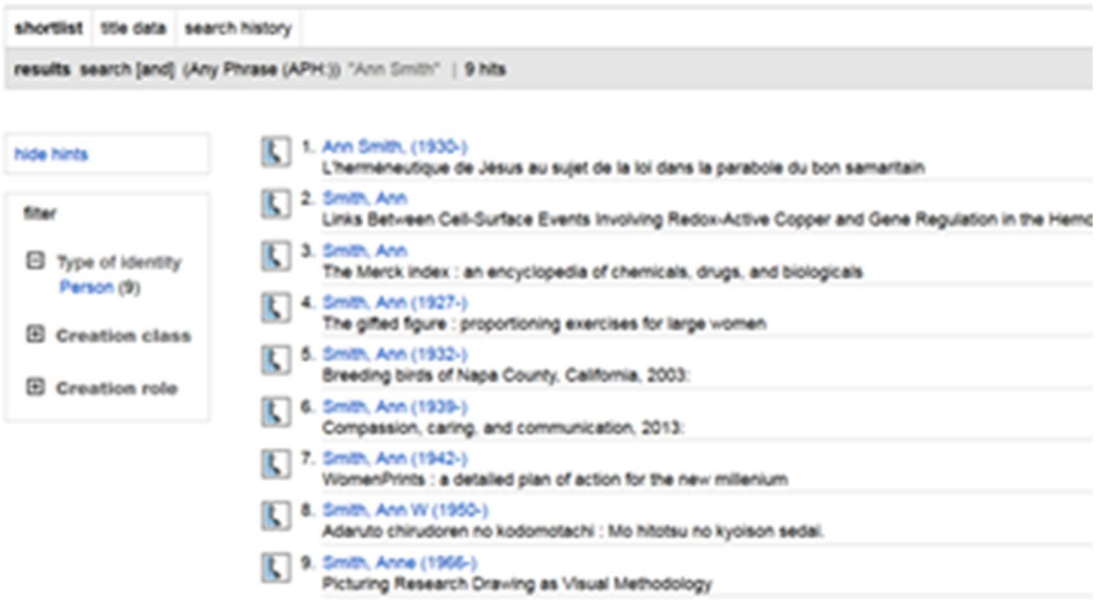


Figure 30: List of ISNI name search results

7. The user can browse the results, and preview candidate ISNI records.
8. The user can modify the search terms if desired.
9. The user locates a record they wish to claim and clicks on it.
10. For the selected ISNI record the user clicks on ‘Add your ORCID’.

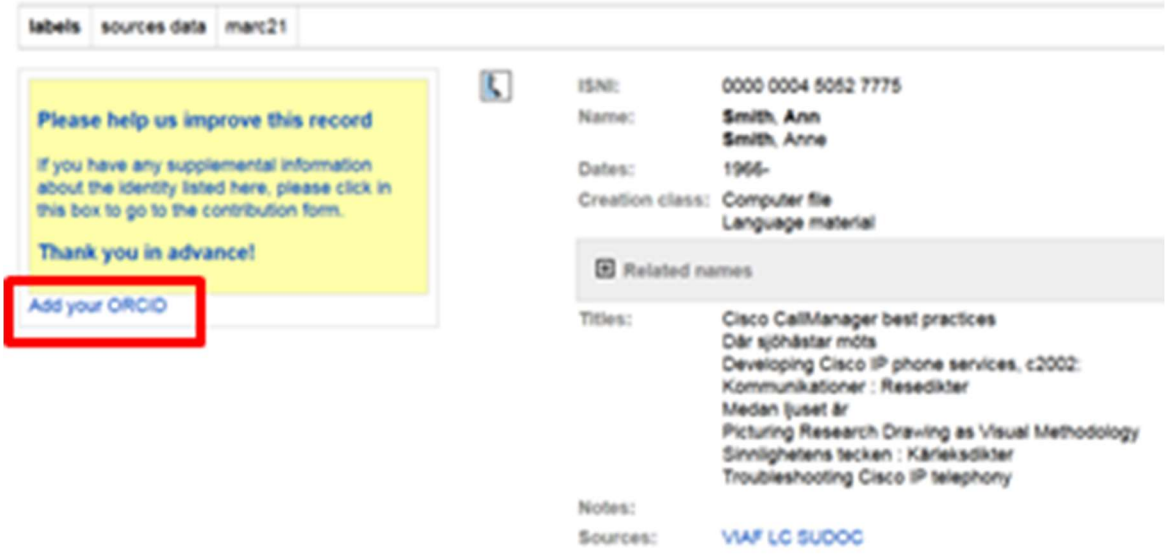


Figure 31: For the selected ISNI record the user clicks on ‘Add your ORCID’

12. The user clicks on ‘Save Changes’. The ORCID iD is added to the ISNI record (red marked field).

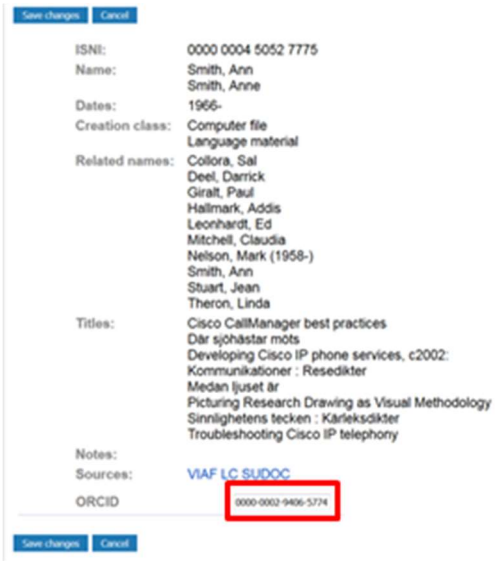


Figure 32: The user clicks on ‘Save Changes’ and the ORCID iD is added to the ISNI record

Requirement R3 states that, as an ISNI user, I would like to see ORCID iDs displayed within ISNI records on the ISNI website if a link is known, so that I can navigate to the associated ORCID record. This is implemented in the following way.

1. The user searches for the individual’s record on the ISNI Public Access Environment by entering a name string or an ISNI ID.
2. The ISNI Public Access Environment displays the ORCID iD within the ISNI record (see Figure 33)
3. The user clicks on the ORCID iD and navigates to the associated ORCID record.

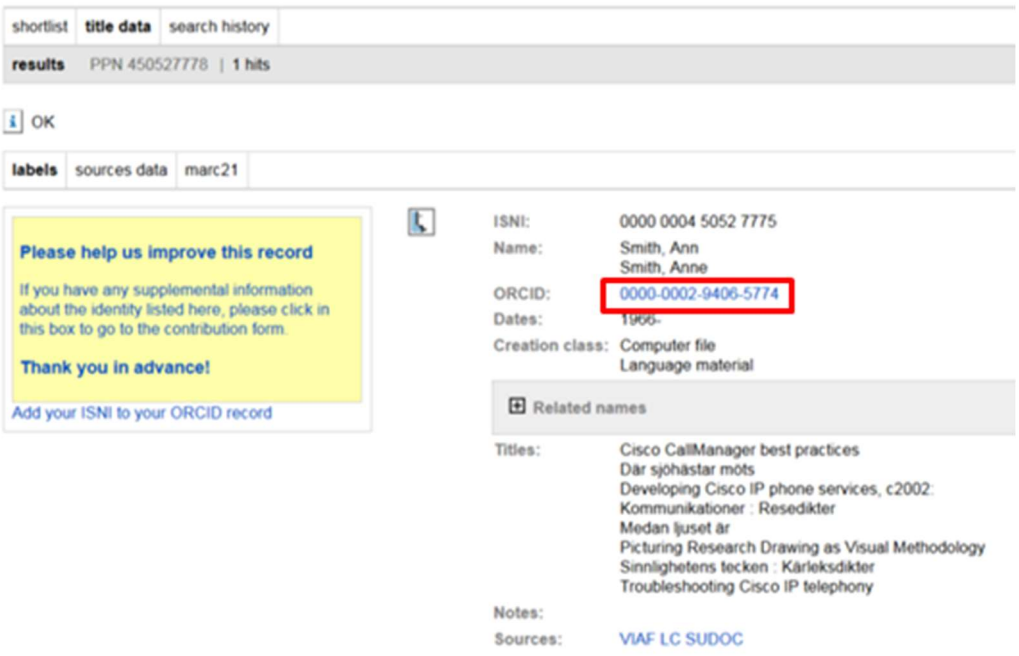


Figure 33: Navigable ORCID iDs displayed within ISNI records

Requirement R4 states that, as an ISNI user, I would like to be able to search the ISNI registry for an ORCID iD so that I can discover which ISNI IDs are linked to it. This is implemented in the following way:

1. The user searches for the individual's record on the ISNI Public Access Environment by entering an ORCID iD in the format "ORCID:0000-0002-0818-6079". It is necessary to explicitly state that the user is searching for an ORCID iD because ORCID and ISNI IDs have the same format and the system cannot know which type of ID is entered.
2. The ISNI Public Access Environment displays the corresponding ISNI record (see Figure 34)

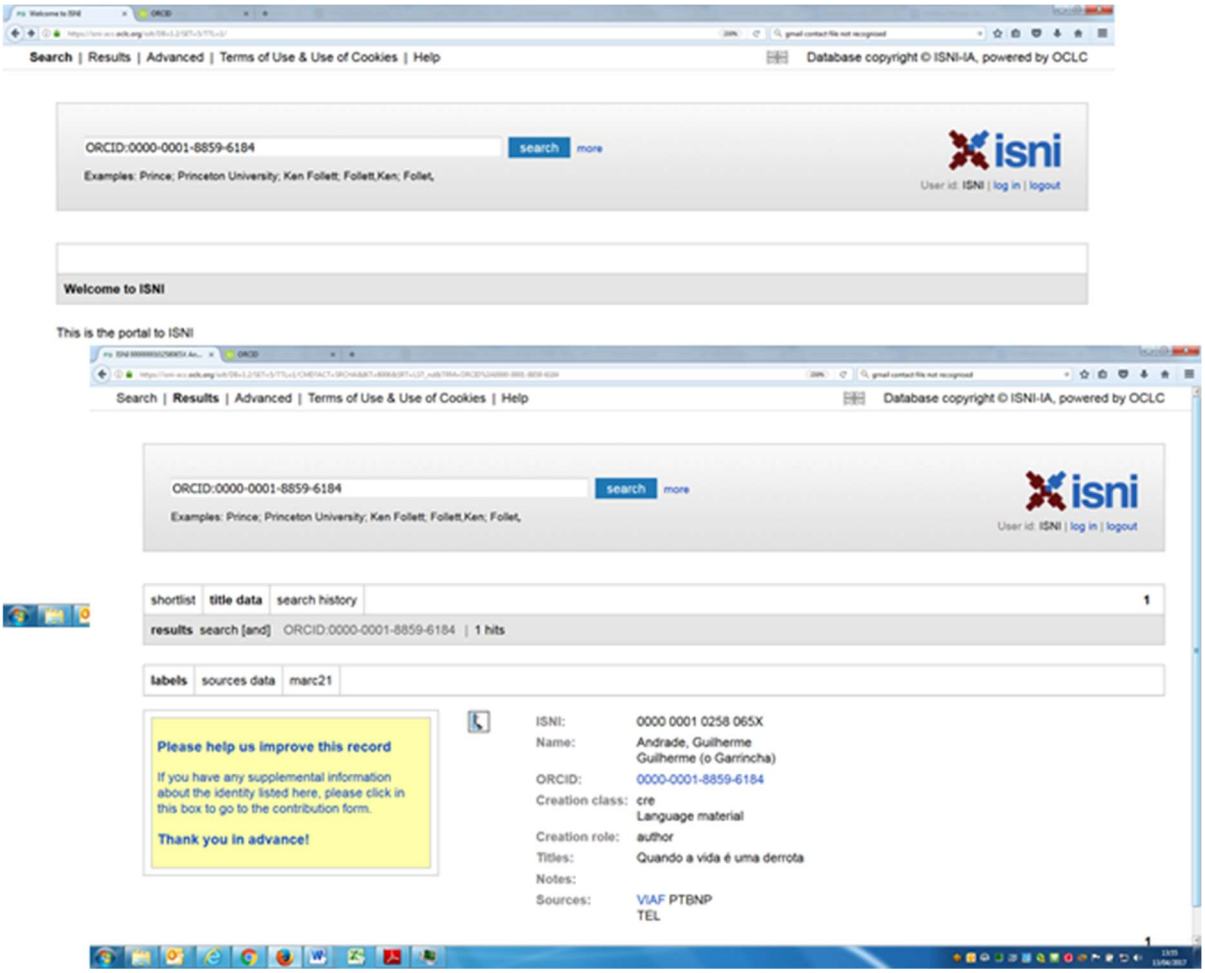


Figure 34: Search the ISNI registry for an ORCID iD to discover the ISNI IDs linked to it

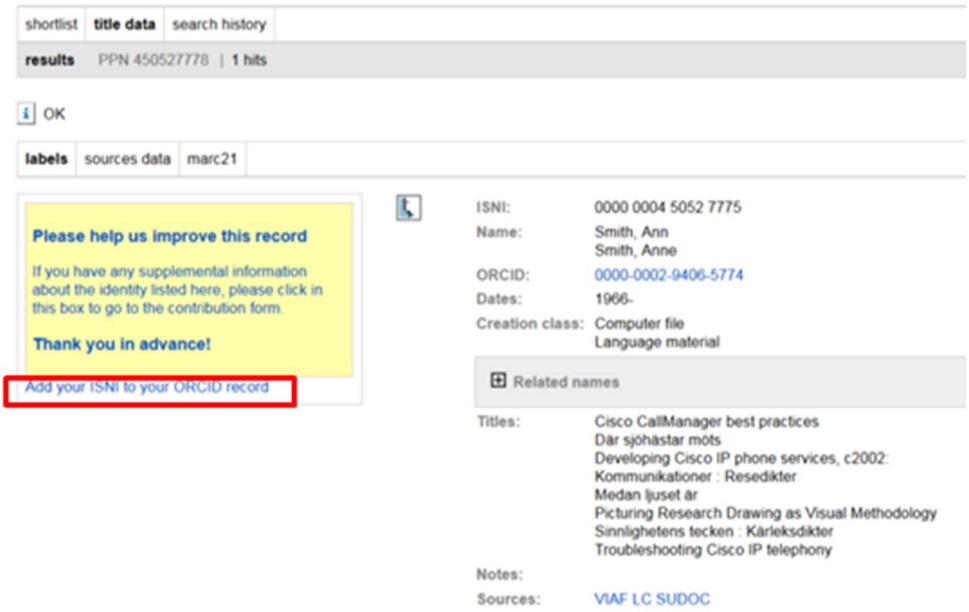


Figure 35: Link to add an ISNI to the user's ORCID record

4.7.2 For ORCID Users

Requirements for ORCID users state that they would like to be able to initiate the linking process to ISNI records from within ORCID so that they can have a more complete scholarly record, and that they would like the user interface to present them with ISNI record results from all name variations after initiating the linking process from ORCID, so that they do not have to manually re-enter search terms and search multiple times. This is implemented in the following way:

1. The user completes steps 1–11, assigning an ORCID iD to an ISNI record from above.
2. The user, having located the ISNI record they wish to claim, clicks 'Add your ISNI to your ORCID record'.
3. The user confirms they are sure. At this point the service updates the user's ORCID record by writing the ISNI identifier to the person record as an external identifier.

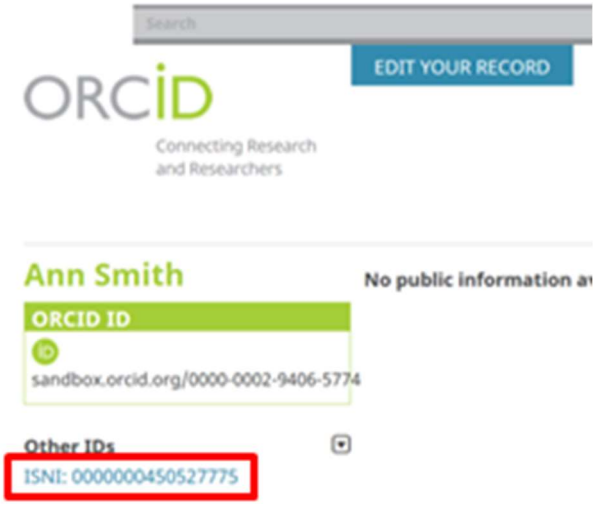


Figure 36: User's ORCID with ISNI added to it

4.7.3 For Developers

As specified in Requirement 7, developers would like to see ORCID iDs included in the metadata attached to an ISNI ID so that they can link the two within their own systems.

1. Having as reference an ISNI Record, such as 0000 0001 0258 065X: (https://isni-acc.oclc.org/xslt/DB=1.2/SET=19/TTL=3/CMD?ACT=SRCHA&IKT=8006&SRT=LST_nd&TRM=0000+0001+0258+065X) (see Figure 37) the developer writes a system that consumes the following URL passing the ISNI 0000 0001 0258 065X as a query parameter: https://isni-acc.oclc.org/sru/DB=1.2/CMD?query=pica.isn+%3D+"0000 0001 0258 065X"&version=1.1&operation=searchRetrieve&stylesheet=https%3A%2F%2Fisni-acc.oclc.org%2Fsru%2FDB%3D1.2%2F%3Fxsl%3DsearchRetrieveResponse&recordSchema=isni-b&maximumRecords=10&startRecord=1&recordPacking=xml&sortKeys=none&x-info-5-mg-requestGroupings=none

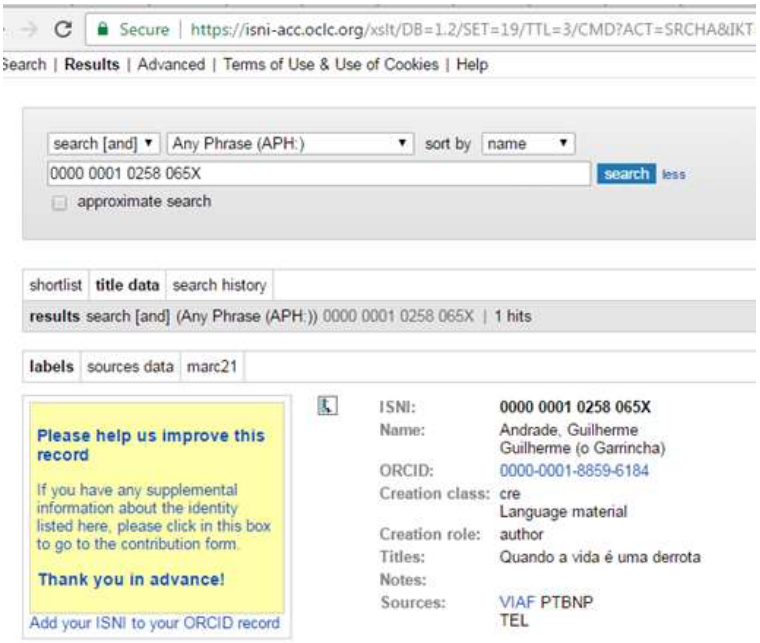


Figure 37: Searching the ISNI tool for an ISNI ID

2. The URL returns XML to the developer, containing the ORCID iD in the metadata (see figure 38)

As specified in Requirement 8, developers would like to be able to search ISNI for an ORCID iD using the ISNI search API so that they can write better software.

1. Having as reference an ORCID iD, such as 0000-0001-8859-6184: (https://isni-acc.oclc.org/xslt/DB=1.2/SET=19/TTL=3/CMD?ACT=SRCHA&IKT=8006&SRT=LST_nd&TRM=00 (see Figure 39), the developer writes a system that consumes the following URL passing the ORCID 0000-0001-8859-6184 as a query parameter: https://isni-acc.oclc.org/sru/DB=1.2/CMD?query=pica.orcid+%3D+%220000-0001-8859-6184%22&version=1.1&operation=searchRetrieve&stylesheet=https%3A%2F%2Fisni-acc.oclc.org%2Fsru%2FDB%3D1.2%2F%3Fxsl%3DsearchRetrieveResponse&recordSchema=isni-b&maximumRecords=10&startRecord=1&recordPacking=xml&sortKeys=none&x-info-5-mg-requestGroupings=none

2. The URL returns XML to the developer, containing the ISNI ID in the metadata (see Figure 38).

Figure 38: ORCID iD is included in the ISNI metadata and the ISNI ID corresponding to an input ORCID iD are returned in XML

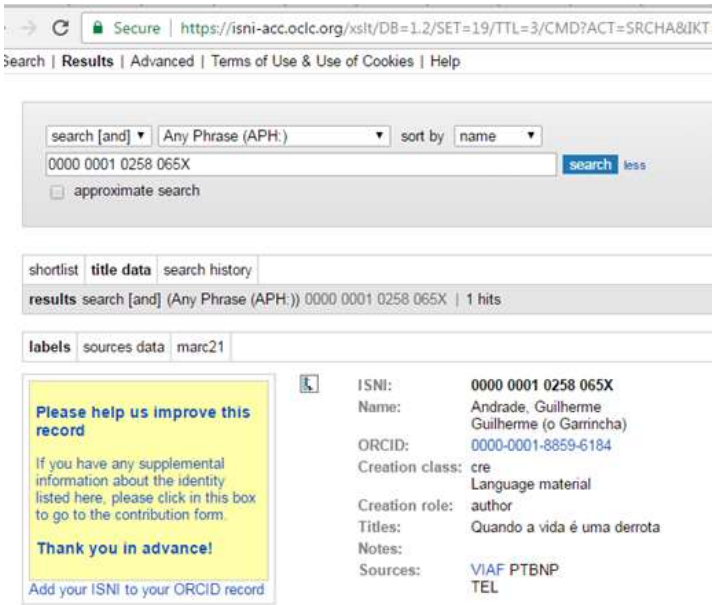


Figure 39: Searching the ISNI tool for an ORCID iD

5 Challenges and Lessons Learned

5.1 ORCID

The cross-linking work outlined in this report is an ongoing effort. Much progress has been made, but as ever there remains much to do.

The workflow for retrospectively linking funding and works is not mature enough for widespread adoption, not least because there is no consensus on where these links should reside. The ORCID facilities working group will continue the exploration of funder-work linking requirements, and hopes to establish a consensus vision of the role ORCID plays in this space.

Where retrospective funding–work links can be stored is still an open question. This question needs to be addressed in concert with the relevant stakeholders: identifier providers, funders, researchers and publishers. In advance of this, the possibility of an expanded vocabulary of relationship types (for example, “funded-by”) is being considered by the ORCID technical team and will be brought up with the API steering group for discussion.

A parallel and significant effort is underway to improve linking and attribution at the point of creation, for example, ensuring that grant identifiers and funding information are captured during the article submission process by publishers, and then exposed alongside article metadata in a machine-readable format. This would reduce or eliminate the need for post-hoc linking, and make the process much more efficient for researchers, funders and systems collecting this information.

ISNI is primarily an identifier service for rights holders. It is not an authoritative source of book metadata. As a result, the tool no longer adds ISBN identifiers or metadata to the ORCID record, it adds only the ISNI identifier itself. This means that a new service will be required to enable researchers find and import ISBN and book metadata for their publications into their record.

Work is continuing on how ORCID deals with the identifier equivalence problem. The technical team is considering strategies for normalising identifier values on ingestion and also in hindsight. The internal report on how ISBNs are constructed within the registry, and how they may be validated and normalised, is being followed up with investigations into other identifiers.

5.2 EMBL-EBI

Situated in the biomedical and life sciences, EMBL-EBI is at the centre of a research landscape of many diverse databases. This means that integrating the whole of biomedical and life science researchers with ORCID, and repository identifiers with DOIs requires the participation of many databases.

While the technical service solution is relatively easy to action, the real challenges come with the human factors involved in repository uptake, since each database acts independently and possesses its own work schedule and priorities. Thus, many conversations are being held in order to align expectations.

At present, the EBI Search integration with batch ORCID claims has gone from the prototyping phase to the implementation, and is currently in alpha version. Thus, the internal work of development continues until the total operationalisation of the system under production is achieved.

In addition to the technical and human challenges of coordinating different groups to work on a common agenda, there are extra costs and risks associated with implementing DOIs across the life sciences databases at EMBL-EBI, which demand thoughtful investigation:

1. A new system of identifiers may be confusing to journals, with which databases have worked for many years to require submissions of data. It will be critical to ensure that the good work already achieved in these fields is not broken, but rather formalised and improved.
2. The reliance on external accessioning and resolving systems holds significant risks for the long-term provision of data services in the public domain, in line with the EMBL-EBI mission.
3. Resolution of non-submission data within EMBL-EBI will still be a requirement that external accessioning systems needs to meet. Therefore for the foreseeable future, DOIs should not replace existing accessioning systems, and the EMBL-EBI will need to maintain a resolving system for all resources.
4. There is hesitation with regard to scalability, considering that worldwide there are around 143 million DOIs (http://www.doi.org/factsheets/DOIKeyFacts.html) handling around 5 billion resolutions per year, including DataCite that has issued around 9.8 million DOIs[4]. Assigning DOIs to all EMBL-EBI data objects overnight would require in the order of 250 million DOIs, posing scalability apprehension. This fact, together with the highly dynamic nature and complex production pipelines of some resources, has led to the assumption that it applies only to resources dealing with submitted 'primary' data, not to value-added, highly-dynamic data resources such as UniProt and Ensembl.
5. There are complexities associated with DOI assignment within data resources produced by international consortia, in which the same data is replicated and available from multiple sites.

5.3 PANGAEA

A key challenge for PANGAEA is the handling of legacy data and metadata that already include identifiers but are not cross-linked according to best practices. The case of IGSNs is a good example. These identifiers have been utilised for some time and are included in data and metadata published by PANGAEA. However, they were not previously curated as they are curated now, due to missing technical requirements (for example, lack of a distinct data type) as well as lagging social adoption (for example, the current relevance of the new PID was not foreseen). In some cases, such as IGSN, it is possible to implement a (semi-)automated process that transitions legacy data to meet state of the art features. However, curatorial steps are typically inevitable. Even a few thousand datasets become prohibitively expensive to process with manual intervention. As a result, a data centre may need to accept the fact that legacy data will not be transitioned to meet state of the art features.

Data centres may want to be cautious in adopting immature PID types. Especially when curatorial processes involve manual steps, data centres rely on maturity and stability regarding PID syntax, representation and resolution. For instance, should PANGAEA represent the IGSN as hdl:10273/CAI000001 or igsn:CAI000001? Should it resolve via https://hdl.handle.net or https://igsn.org? While it is not too much of a hassle to exchange one for the other, it is arguably also true that the former representation

[4] http://stats.datacite.org

and resolution service may be more stable than the latter. While for DOI the answer is clear, for IGSN it is not.

Perhaps the more important lesson learned for PANGAEA is that, contrary to ORCID integration, cross-linking is a continued effort, one that involves an expanding number and type of more or less mature PIDs. The work on cross-linking is thus not completed but will continue.

5.4 CERN

While implementing cross-linking between CERN Analysis Preservation and resources specific to the experimental collaborations, we encountered challenges, largely of a social or political nature.

As alluded to previously in this document, it is important to note that CERN collaborations operate independently. This is not a by-product of the scale of these collaborations, as it may appear to the outside, but rather a key component of ensuring the accuracy and replicability of scientific results. The experimental equipment and platforms the collaborations use (namely, the LHC and each component apparatus, such as ATLAS or CMS) are unique, meaning that independent verification of results by other labs is essentially impossible. If only one group was producing results, or if only a single experiment was being run, it would be difficult to ascertain whether or not these results are accurate. In order to create an environment where verification is possible, multiple experiments are designed and subsequently run on the LHC. Only after careful analysis within the collaborations are the results shared across the HEP community for wider confirmation and expansion.

This means that there are understandable concerns over allowing a service external to the collaboration unfettered access to a collaboration-specific database. The CERN Analysis Preservation team has therefore worked hard to liaise with the collaborations and to involve them in every step of the way in the database-linking process. As described in the requirement section of this document, it was important to keep the analyses separate and to ensure that specific collaborations and/or working groups have complete control over the confidentiality and sharing of their analyses. In practice, the separate service account creation described in the requirements section was one step taken to reassure the collaborations that we understand and are sensitive to these access concerns. Connecting to collaboration-specific information resources still requires a great deal of trust achieved through consultation, discussion, and demonstration with the individual collaborations. The results of these ongoing conversations are evident in the current slate of collaboration-specific databases to which CERN Analysis Preservation connects, as well as in the flexible custom JSON schemas that structure the relevant collected data.

5.5 DataCite

As mentioned in the previous section, there are a number of additional requirements that DataCite will need to address in the future in order to take this prototype forward.

A challenging requirement facing DataCite's services is that retrospectively adding information to authors' works must not overwrite the funding information initially provided by DataCite's data centres. This means that if a work was associated with a funder by the data centre, authors would not be able to modify the relationship. This is currently a prerogative DataCite grants to the information coming from data centres for all other services that retrospectively add information to the works in DataCite (for

example, ORCID claiming). Therefore, DataCite needs to build the infrastructure to deal with an extra layer of storage for this information to be managed.

Another requirement that DataCite's services face is the need to accommodate the differences in the metadata schema versions associated with works. The DataCite metadata schema has been evolving throughout the years, and thus has the way in which funding or funder were modelled in the schema. In earlier versions, funder was modelled as a type of contributor (for example, "contributorType") within the contributor property. In its latest version, funder is modelled as a property in its own right, using the "FundingReference" property. These differences and the transformation between the current version and previous ones need to be addressed.

An additional requirement is the ability to push works-to-funders relationships to authors' ORCID record. The challenge here is twofold. First, we need to extend the DataCite EventData service to cope with this relationship. Second, we need to coordinate with ORCID around the way in which the relationships would be stored in their schema and vice-versa. We have initiated such coordination but further work needs to be carried out. The prototypes presented by DataCite and ORCID in this deliverable have demonstrated the way authors deal with linking funding information to works and persons; the next challenge is to connect the prototypes to make it possible.

5.6 British Library

The ISNI–ORCID integration tool needs to be developed further in order to create a smoother experience for users. At present, the tool is a closed alpha version, and needs some additional work to meet the requirements set out in the original tender (as described in section 3.6 above).

The ISNI and ORCID approaches vary significantly in the following ways. ORCID iDs are researcher identifiers with metadata curated by active researchers (or their delegates), by trusted linked systems that link works through 'Search and Link Wizards', or by trusted organisations via an authenticated API connection. ORCID iDs are explicitly designed to be owned and controlled by the individual to whom they refer, and they are built to be incorporated into workflows and therefore generate unambiguous links between individuals and contributions to research (for example, reviews, funding or works) at the point of creation. Trust issues are managed by exposing detailed information about the history and provenance of an assertion. ISNI identifiers and their metadata are assigned by an authority and assured by a quality team. They are primarily used in the cultural heritage and rights management communities, rather than in research administration and scholarly communications. They go beyond the scholarly record and can include any named entity (including researchers both living and historical, organisations and fictional characters). Because of this, the ISNI system is not designed to bring out the fact that an assertion, such as an ORCID iD, has not been made by the ISNI quality team. Further work is needed to make the source of the metadata explicit.

Additionally, since the ISNI system is usually edited by professional library staff, the information that is presented to ORCID users may be overwhelming. More work has to be done to present the users with a simpler landing environment once they enter the ISNI system.

ORCID 'Search and Link Wizards' normally link to other sources of works as well as person identifiers. There is a prominent 'Add Works' button that initiates the linking between these systems. The new version of the ISNI-ORCID link tool, however, only creates person identifier links, which clients can

use to infer person–work links by using the ISNI search API. As such, the developed tool is unique in its approach and can only be incorporated into the ORCID User Interface if a feature for adding person identifiers without works is added at the ORCID end.

There is a risk that ORCID identifiers linked to an ISNI record are removed at the ORCID end but not updated at the ISNI end (due to user error or because they've been misassigned). The usual workflow is for users to go through the search and link wizard to remove the links. However, there is still the possibility that the link is removed by the user at the ORCID end without invoking the wizard. In these cases, how ISNI could periodically check to see if links are still valid is an open question. Adding an extra confirmation step to the wizard would hopefully reduce erroneous assertions.

6 Conclusion

THOR participants comprise a wide range of disciplinary areas such as the biological and medical sciences, environmental and earth sciences, physical sciences, social sciences and humanities. Several independent services have been developed by the individual THOR partners in order to realise the common goal of delivering solutions for the integration of multi-disciplinary PID services to the research community.

Each of the linking tools or service enhancements created by the project partners has improved the usage and utility of both the PIDs employed, and the service provided to researchers. However, some challenges have been encountered in this process, and it is interesting to note that these can be shared across disciplinary areas.

Retrospective claiming and the creation of PIDs for existing resources pose real challenges of scale and effort. For many services and communities, it may be necessary to keep historical resources in their current state (for now) and focus on ensuring that new accessions are assigned PIDs going forward.

Trust in newly created PID systems, or even newly adopted pre-existing PID systems, takes time to build. Political and social changes take time, consensus-building, and communication to take effect. This is the role of the communication work package within THOR; this report provides a useful source of know-how and evidence for that team to exploit and promote.

While technical details vary across systems, and requirements vary across communities, the challenges lie in common areas – and these areas are accurately described as social. Adoption and integration rely on organisations to change the services they offer and individuals to change the way they work. Reducing these barriers is truly the work of the whole community. The efforts described in this document help to demonstrate that the benefits of PIDs are real and within reach. This is surely the first step toward delivering the consensus necessary to create change.

7 References

de Mello, G., Graef, F., Stocker, M., Schindler, U., Dasler, R., McEntyre, J., & Dallmeier-Tiessen, S. (2016). Demonstration of Services to Integrate ORCIDs into Data Records and Database Systems. Zenodo. https://doi.org/10.5281/zenodo.58971

ODIN Consortium, Amir Aryani, Amy J Barton, jan brase, Josh Brown, Tom Demeranville, Patricia Herterich, Lynne McAvoy, Laura Paglione, Sergio Ruiz, Gudmundur Thorisson, Todd Vision, Frauke Ziedorn(2015). D4.2: Workflow for interoperability. Figshare 10.6084/m9.figshare.1373669.v1

Basic tutorial: Webhooks notifications. (2017). Retrieved from http://members.orcid.org/api/tutorial/webhooks

Appendix A: Project Summary

The **THOR** project will 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 project has four concrete aims:

1. Establishing interoperability
2. Integrating services
3. Building capacity
4. Achieving sustainability

The project will meet these aims by defining relations between contributors, research artefacts (including data), and organisations. We will incorporate these relationships into the ORCID and DataCite systems. We will also expand existing linkages between different types of identifiers and versions of artefacts to improve interoperability across platforms and integrate ORCID iDs into production systems for article and data submission services in pilot communities and beyond.

The consortium will develop systems to embed new PID resolution techniques into existing services to support seamless direct access to artefacts, and in particular data. We will create services to allow associations between datasets, articles, contributors and organisations at the time of submission. Building on these, we will deliver the means to integrate trans-disciplinary PID services in community-specific platforms, focussing on cross-linking, claiming mechanisms and data citation (guided by the FORCE 11 data citation principles).

For more information, visit http://project-thor.eu or contact mailto:info@project-thor.eu.

Appendix B: Terminology

Additional terms are defined below:

Term	Definition
API	Application Programming Interface
ATLAS	One of the four major experiments at the Large Hadron Collider at CERN
CERN-SIS	CERN Scientific Information Service
CIENCIA-IUL	Identifier used by the ISCTE-IUL CRIS system for national reporting in Italy
CMS	Content Management System
CRIS	Current research information system
DataCite	An organisation that develops and supports methods to locate, identify and cite data and other research objects. Specifically, DataCite develops and supports the standards behind persistent identifiers for data, and the members assign them. See https://www.datacite.org
DOI	Digital Object Identifier
EFO	Experimental Factor Ontology
ENA	European Nucleotide Archive
HEP	High Energy Physics
ID	Identifier
IGSN	International Geo Sample Number
ISNI	International Standard Name Identifier
ISO	International Organization for Standardization
ISSN	International Standard Serial Number
JISC	Joint Information Systems Committee, a United Kingdom not-for-profit company
KUID	KoreaMed Unique Identifier
LENSID	LENS Patent identifiers
LHC	the Large Hadron Collider
ORCID	An organisation that creates and maintains a registry of unique researcher identifiers and a transparent method of linking research activities and outputs to these identifiers. See http://orcid.org
PDB	Protein Data Bank in Europe
PID	Persistent Identifier
PMC	Pub Med Central
RDA	Research Data Alliance
RDF	Resource Description Framework
RINGGOLD	Customer identifier
UI	User Interface
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
XML	Extensible Markup Language