Services that Support Claiming of Datasets in Multiple Workflows

Document Information

Date: 13/02/2017

Authors: Guilherme de Mello (EMBL-EBI), Florian Graef (EMBL-EBI), Markus Stocker (PANGAEA), Uwe Schindler (PANGAEA), Robin Dasler (CERN), Johanna McEntyre (EMBL-EBI), Kristian Garza (DataCite), Sünje Dallmeier-Tiessen (CERN)

Reviewers: Angela Dappert (BL), Laura Rueda (DataCite)

Abstract: This report summarises progress on enabling researchers and other contributors to associate datasets with their ORCID record. This is an important advance in enabling unambiguous attribution and credit for research. 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.290649

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 November 2016.

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. They 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.

This report summarises progress on enabling researchers and other contributors to associate research artefacts with their ORCID record, a process known as *claiming*. The dataset claiming process involves creating, maintaining, and sharing information about the relationship between researchers and datasets.

We describe our experience implementing the claiming process at three different organisations in three different disciplines in three different countries. We identify some of the shared challenges as well as the unique issues each organisation faced developing and successfully deploying the claiming process into a live operational production system.

This is an important advance in enabling unambiguous attribution and credit for research.

While technical challenges remain, such as synchronisation of claims, technical capabilities have substantially improved. The human and social challenges are now coming to the fore – we must ensure that claiming services are widely adopted and used across the research communities.

Contents

1 Introduction

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. They 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.

This report summarises progress on enabling researchers and other contributors to associate research artefacts with their ORCID record, a process known as *claiming*. The dataset claiming process involves creating, maintaining, and sharing information about the relationship between researchers and datasets.

ORCID maintains a registry of unique researcher identifiers and links from these identifiers to research activities and artefacts. Integrating ORCID in disciplinary data repositories enables automated attribution of the contributions researchers make. This lays the foundation for securing recognition and credit for research outputs beyond traditional article publications, specifically tasks such as collection of data, creation, documentation and publication of datasets.

In previous work (de Mello et al., 2016), we reported on successful integrations of ORCID identifiers into disciplinary data repositories. This work enabled the repositories either to use ORCID iDs as a primary identifier for user accounts, or to associate an ORCID iD with a user account. This report describes the next steps, in which we extended the implementations to support a full claiming workflow with a high level of automation.

Generally a claim is made by the contributor for a given dataset, and the claim results are updated according to the ORCID record. In general, a claim can be defined as: to state or assert that something is the case, typically without providing evidence. For instance, a tourist claims to have visited the Eiffel Tower in Paris. In the scientific milieu, a claim could be when a postgraduate student claims to be the author of an article, or a researcher claims to have contributed to a dataset. It is important to note that a claim can be performed at the time of submission (pre-publication) or retroactively (post-publication), where either the researcher or the data centre sends the claim information to ORCID.

In the case of a disciplinary data centre, a claim may be made at any time relative to the time that a research artefact becomes available or published. It is common for claims to be made at the time of submission (pre-publication) or retroactively (post-publication). Either the researcher or the data centre sends information about the claim to ORCID. ORCID receives the claim information and ensures that it has permission from the relevant researcher to record or update the information in their ORCID record and to display it. Figure 1 provides a schematic overview of claiming. Because ORCID may receive claim information from many data centres, it can provide a joined-up view of both a researcher's outputs and the set of researchers associated with a single output.

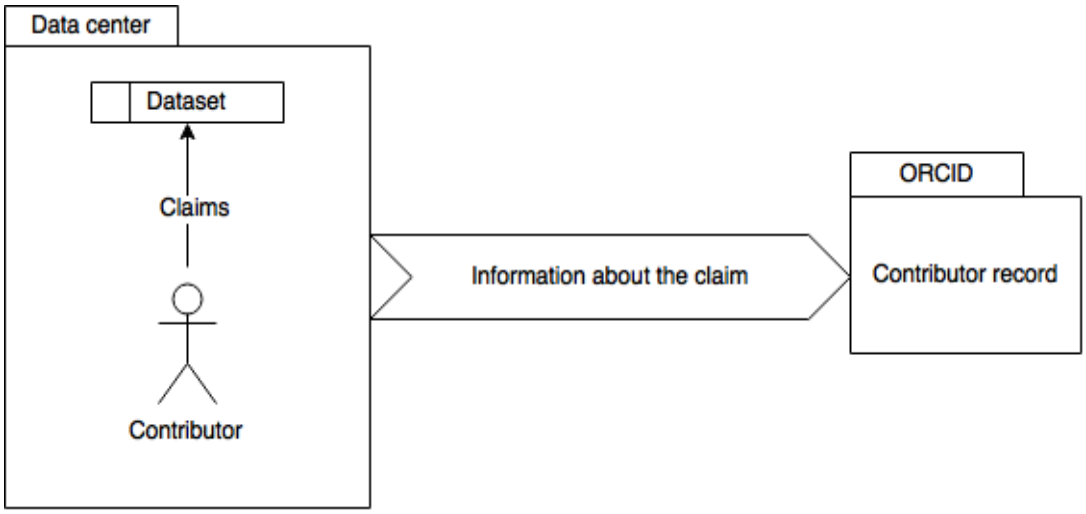

Figure 1: Schematic overview of the claiming process.

In order to link contributors and datasets, claiming services rely on the integration of ORCID iDs. Thus, data centres need first to obtain and maintain the (validated) ORCID iD of contributors. They also need to communicate information about the relationship between ORCID iDs and dataset DOIs to ORCID, so that the ORCID record of researchers can be updated accordingly.

This report describes how EMBL-EBI, PANGAEA, CERN, DataCite, and ORCID have collaborated to develop services that push information about these relationships to ORCID. The three data centres – EMBL-EBI, PANGAEA, and CERN – serve three distinct disciplines, namely Life Sciences, Earth and Environmental Sciences, and High-Energy Physics, and operate from three different countries, namely the UK, Germany, and Switzerland, respectively. The data centres therefore operate in different environments, with different legacy workflows and technologies. Being a PID infrastructure, DataCite is an altogether different kind of system compared to data centres. To be successful, dataset claiming implementations need to respect these differences, considering the distinct requirements carefully as well as building on employed technologies and integrating into the existing workflows.

Though the four institutions share a common goal, the differences at each institution entailed different solutions to achieve it, leading to complementary results:

- EMBL-EBI hosts several independently governed and separately funded databases, necessitating a centralised service that pushes claims directly to ORCID.
- PANGAEA is not an ORCID member and thus pushes claims for data publication collections to DataCite as a way to update ORCID records.
- CERN's main goal is to eliminate any need for retrospective claiming, focusing instead on automating metadata collection up front.
- DataCite, with its dataset-centric view, implements dataset self-claiming at individual publication level – an approach pursued also by EMBL-EBI but not by PANGAEA.

We start by describing the institutions and provide a brief overview of their services. We then present dataset claiming requirements, followed by a presentation of the work completed and the results obtained from the perspective of each institution. Finally, we discuss the shared challenges.

2 Institutions and Services

Several THOR partners were involved in developing the dataset claiming services discussed above. To provide readers with the context in which the claiming services were developed, in this section we briefly present the institutions involved and their services: first the PID infrastructures, ORCID and DataCite, and then the disciplinary repositories, EMBL-EBI, PANGAEA, and CERN.

2.1 Service Providers

2.1.1 ORCID: Open Researcher and Contributor ID

ORCID[1] 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 an ORCID identifier, manage their record of activities, and search for others in the registry. Organisations are required to become members to link their records to ORCID identifiers, to update ORCID records, to receive updates from ORCID, and to encourage their employees and students to register for ORCID identifiers via the create-on-demand process. All public data are made freely available via periodic data dumps and the API.

The ORCID API allows systems and applications to connect to the ORCID registry, including 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.

2.1.2 DataCite: 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. This interface complements the DataCite OAI-PMH service, providing an actionable endpoint to query DataCite's metadata collection.

[1] http://orcid.org/

- **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 documentation. Event Data incorporates software agents that look for DataCite DOI links in specific parts of the web (such as Crossref and GitHub). This service is now also used to keep track of dataset claiming events made by users.
- **DataCite Content Resolution**, intended for the technical side of dataset management, 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.

In 2013, DataCite – as part of the EC-funded ODIN project[2] – also launched a pilot service that allowed users to claim DataCite DOIs to their ORCID record[3]. Since 2015 this functionality is part of the standard DataCite Search service, **DataCite Search and Link**. The DataCite Search and Link service allows contributors to search DataCite metadata via the DataCite Search interface, and to claim DOIs and metadata to their ORCID record. In essence, 'search and link' functionality depends on authors claiming datasets after their publication. Every time an author claims a DOI in the DataCite Search and Link service, the author's ORCID record is updated. The functionality has been extended to include ORCID auto-update: if the researcher allows the connection, every time a DataCite DOI is minted, his or her ORCID record will receive an automatic update.

2.2 Disciplinary Data Repositories

2.2.1 EMBL-EBI: The European Bioinformatics Institute

The European Bioinformatics Institute (EMBL-EBI)[4], a centre for research and services in bioinformatics, is 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)*

[2] http://cordis.europa.eu/project/rcn/105189_en.html

[3] http://dx.doi.org/10.6084/m9.figshare.824314

[4] http://www.ebi.ac.uk/

In the life sciences, it is uncommon to use DOIs as (primary) persistent identifiers. Most repositories assign their own identifiers to their data records. This practice has led to hundreds of identifier types – almost every database has its own. Some databases complement their own identifiers by assigning DOIs to their records as a secondary identifier. The self-assigned identifier numbers, however, are not resolvable to a data record on their own, but the knowledge of the identifier type is required to access the data.

The identifiers.org service aims to solve this issue by resolving to the correct data record when identifier type and accession are given, so that stable URLs can be formed.

The plethora of identifier types, such as ArrayExpress, BioModels, ChEBI, Ensembl, ENA, EFO, Expression Atlas, Gene ontology, InterPro, MetaboLights, PDB, UniProt, and others, poses a scalability problem for ORCID. A solution based on the identifiers.org service is in development but, in the meantime, the 'other' identifier type is being used as a container holding a concatenation of identifier type and identifier number.

The Literature Services team at EMBL-EBI maintains and develops Europe PMC (Europe PMC Consortium, 2015)[5] – a search engine for publications in the biomedical life sciences. As part of Europe PMC's service, the ORCID iD system is used to uniquely identify authors in literature searches. Authors are also able to claim any publications in Europe PMC to their ORCID profile directly on Europe PMC. This integration with ORCID, together with continued support for ORCID adoption and an EMBL-wide adoption of ORCID iDs (including the infrastructure and services to assign staff-ORCID iDs), has led to 3.3 million articles in Europe PMC now having links to ORCID iDs.

Many essential life science services that are run at the EBI could benefit from ORCID integration. Through the experience of integrating with ORCID, the Literature Services team is in the right place to reach out and increase ORCID adoption, and improve data–literature interlinking. This deliverable documents how we have built on de Mello et al. (2016) at EBI, and extended our central ORCID gateway with functionality that not only authenticates and gathers information from ORCID profiles but allows the claiming of data records to ORCID.

2.2.2 PANGAEA: Data Publisher for Earth & Environmental Science

PANGAEA, the Data Publisher for Earth & Environmental Science[6], is an information system operated 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 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.

[5] http://europepmc.org/

[6] http://www.pangaea.de/

Data publication claiming is enabled by the recent THOR-facilitated ORCID integration in PANGAEA (de Mello et al., 2016). With the integration, contributors to data publications can now provide PANGAEA with their validated ORCID iD. Creating this link is sufficient for users and contributors to claim all their datasets deposited at PANGAEA. PANGAEA automatically pushes the claims to DataCite. The claims then appear on the contributor's ORCID record automatically via ORCID Auto-Update.

An important difference between PANGAEA and other systems, including those described in this document, is that by creating the link between their PANGAEA user profile and their ORCID iD, contributors instantly claim *all* their datasets. This approach stands in contrast to systems where users claim publications individually.

2.2.3 CERN: European Organization for Nuclear Research

CERN[7], the European Organization for Nuclear Research, maintains a variety of services for managing both literature and data outputs of High-Energy Physics (HEP) research. CERN Scientific Information Service (SIS), the THOR partner, is specifically involved in the following:

- **INSPIRE**[8,9] – Our core service, a hub for literature in HEP. Built as an aggregator, it automatically harvests relevant literature from a list of publishers. Literature relevancy is determined largely through automation with a human approval and suggestion layer.
- **HEPData**[10,11] – A collaboration between CERN and Durham University explicitly for data that are supplemental to publications (for example, final tables); includes peer review workflows so that submitted data may be approved prior to release.
- **CERN Analysis Preservation**[12,13] – An internal tool for capturing the disparate files, metadata, and provenance associated with large-scale HEP analyses.

INSPIRE provides author profiles. These profiles are generated automatically from the harvested papers by a clustering algorithm that operates behind the scenes. In order to improve the effectiveness of the clustering and correct errors, users are provided the opportunity to claim papers to their INSPIRE author profile. Any papers associated (through clustering or claiming) with an INSPIRE author profile that has an ORCID iD attached are appended to that user's ORCID record as part of regular pushes.

For the purposes of demonstrating retrospective claiming in HEP systems, we have extended this claiming function to datasets on INSPIRE Labs, the INSPIRE new feature showcase. INSPIRE is currently undergoing significant new development as we upgrade its underlying platform and implement newly possible features. As new features are developed for INSPIRE they are introduced on INSPIRE Labs, while legacy INSPIRE still serves as the primary entry point for users. Just like the paper claiming that is currently being ported to INSPIRE Labs, dataset claiming will allow datasets to be appended to a user's ORCID record. Additional internal discussion and outreach work is warranted before this feature is rolled out wholesale to all users, but the work described here demonstrates the feasibility of providing a dataset claiming option to HEP users.

---

[7] https://home.cern/

[8] http://inspirehep.net/

[9] https://github.com/inspirehep

[10] https://hepdata.net/

[11] https://github.com/HEPData

[12] https://analysis-preservation-qa.cern.ch

[13] https://github.com/cernanalysispreservation

3 Requirements for Claiming Datasets

This section focuses on the key requirements for implementing dataset claiming at the three disciplinary repositories and DataCite. The requirements highlight the key differences among the institutions, their services and workflows, including:

- At EMBL-EBI:
  - To implement a mechanism in the EMBL-EBI tool that allows researchers to claim to their ORCID profile the authorships and participation on work studies and datasets available on the EMBL-EBI database.
- At PANGAEA:
  - To implement a mechanism in the PANGAEA repository that allows authors to claim all of their datasets instantly, rather than individually.
  - To implement a mechanism in the PANGAEA repository that pushes claim information to authors' ORCID profiles indirectly via DataCite.
- At CERN:
  - To provide a claiming interface for datasets in INSPIRE.
  - To provide a mechanism in INSPIRE that pushes claim information to authors' ORCID profiles, as for other works.
- At DataCite:
  - To provide a mechanism for DataCite Search and API services that pushes claim information to data centres.

3.1 EMBL-EBI

To address the requirement for Biological and Medical Sciences claims, both across multiple databases and within suitable workflows, the EMBL-EBI tool to claim datasets to ORCID is built on the current service for ORCID authentication. As described in de Mello et al. (2016), this ORCID authentication service was developed as a middleware layer (now called the EBI ORCID Hub), which includes an API library to be incorporated into the application, client side. This enables EMBL-EBI databases to seamlessly incorporate ORCID IDs into their resources. The dataset claiming tool consists of interfaces providing high level APIs. This permits the integration with diverse client applications by masking the heterogeneity of processes and simplifying the access to the user's ORCID profile.

A central ORCID linking service ensures that data resources at the EMBL-EBI are integrated with ORCID, and there is no redundancy of effort (for example, duplicate work by two or more databases), making this single service a cost-effective design decision. This core data claiming service has also now been extended with a feature called retrospective claiming, which applies to existing research data and metadata in science repositories. Retrospective claiming allows researchers to claim into their ORCID profile the work studies and datasets available on EMBL-EBI databases that they have previously authored or contributed to.

3.2 PANGAEA

There were two key requirements for the implementation of data claiming at PANGAEA:

1. Authors instantly claim all their datasets, rather than individually
2. To push claim information to authors' ORCID profiles indirectly via DataCite

The first requirement is driven by how PANGAEA manages system users, data publication contributors, and data publications as well as their relationships. PANGAEA manages two types of accounts: contributor and user. Any author named in at least one data publication has a **contributor account**. These accounts are created by PANGAEA data curators. Contributors may have a **user account**. These accounts can be created directly by contributors. Contributors who have submitted data for publication must have a user account.

User accounts and contributor accounts can be linked. This relationship is established manually by PANGAEA curators. PANGAEA maintains associations to ORCID iDs on both user accounts and contributor accounts; for linked accounts, the associations to ORCID iD are synchronised.

Contributors with an established association between their contributor account and their ORCID iD have effectively claimed all their data publications. It is only the contributors without this association that need to claim data publications. Simply establishing this association is sufficient to claim all data publications by the contributor: rather than claiming individual data publications, at PANGAEA one claims the entire set of data publications by 'claiming' the contributor account.

The second requirement is driven by how PANGAEA communicates with the wider network of PID infrastructures. DataCite is the primary hub through which PANGAEA shares metadata about data publications with PID infrastructures. Hence, PANGAEA submits metadata about the relationships between contributor ORCID iDs and data publication DOIs to DataCite. DataCite then distributes this relationship to relevant PID infrastructures, specifically to ORCID.

This approach comes with the primary advantage that PANGAEA shares information only with DataCite, rather than with both DataCite and ORCID. We thus implement fewer interfaces. Moreover, as PANGAEA is currently not an ORCID member, it cannot directly push information to ORCID. The approach also addresses this limitation.

3.3 CERN

The development of information services at CERN follows a philosophy of relieving end users of the burden of metadata provision by automating as many processes as possible, and by establishing a provenance chain up front. Because of this, the primary data service, HEPData, requires an existing paper to be associated with a dataset at the point in which a dataset record is created. The metadata for this paper, including author information, is then linked from INSPIRE, the primary publication record service. This means that HEPData stores no bibliographic metadata of its own and that the authors of the paper are automatically considered the authors of the dataset. In so doing, CERN are seeking to eliminate the need for after-the-fact data claiming altogether. However, HEPData DOIs and metadata were not being pushed to authors' ORCID profiles.

As described in de Mello et al. (2016), a claiming function already exists in INSPIRE, but its use is limited to publications. This claiming function suggests papers to a logged-in user that the system believes to be his/hers. Users can then verify the list of suggested papers, and claim the appropriate papers to their profile individually. There were then two requirements for the demonstration of data claiming at CERN:

1. Provide a claiming interface similar to that for publications
2. Push claim information to authors' ORCID profiles, as for other works in INSPIRE

The implementation for these requirements is described in more detail in the following sections.

3.4 DataCite

One important limitation of DataCite Search and Link is that claims are sent from DataCite Search directly to the ORCID registry. Hence, the data centre hosting the claimed dataset is not informed about this new link between the dataset DOI and the ORCID record. In this way, the DataCite DOI metadata is not updated, as only the original data centre is in control of its content.

Data centres can fetch this information from the ORCID registry using the ORCID API, but there is no specific API endpoint that shows all claims by data centres. Data centres that are ORCID members can take advantage of the ORCID notifications service; but this service is currently built around notifications when a specific ORCID record is updated, not when a DOI from a specific data centre has been claimed.

DataCite data centres have great interest in a mechanism that enables notifications about datasets claimed in the DataCite Search and Link service. The mechanism should allow them to associate already published datasets with ORCID iDs. It would also be interesting for data centres that have not yet implemented the linking to ORCID accounts on data submission. This need is DataCite's main requirement: to provide a mechanism that pushes data claims to data centres. In section 4.4, we show the implementation of a mechanism that aggregates dataset claims by data centres and the interfaces DataCite stakeholders can use to benefit from this mechanism to address their needs.

4 Work Completed

4.1 EMBL-EBI

Facilitated by THOR, the EMBL-EBI has been operating a middleware layer called the 'EBI ORCID Hub', which offers user authentication services for EBI databases via ORCID. The platform has now been improved to incorporate new features such as dataset claiming to ORCID, and new queries that EBI databases can use to harvest ORCID information about the user's works and datasets.

The middleware mediates between software, unifying information exchange between different applications and environments. It is used to move information between EBI databases and ORCID, simplifying the differences in platforms, programming languages, and communication protocols.

The EBI databases are data repositories hosted at EMBL-EBI that provide a wide range of free data and distinct types of technological applications to support various life science studies, enabling researchers to find, publish, and share experimental findings. Currently, these databases hold thousands of published works and datasets, along with detailed information about them in community-agreed standards. The integration of these EBI databases (Figure 2) with the EBI ORCID Hub allows for submitted data to be claimed by the researchers who contributed to the corresponding datasets. Information about claimed datasets is then automatically added to the contributor's ORCID record, making it accessible to the scientific community.

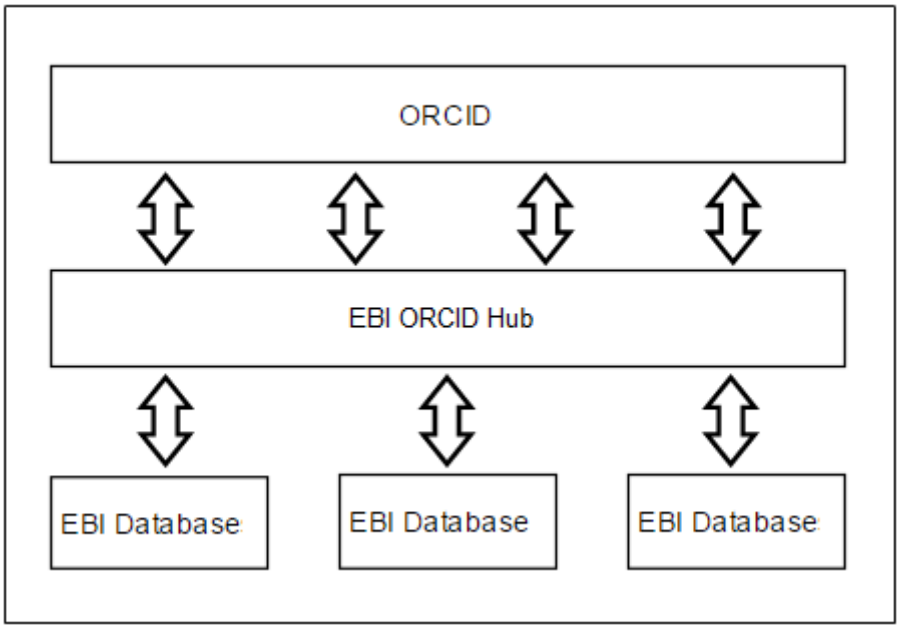

Figure 2: Overview of EBI databases accessing ORCID through middleware service

The steps required to claim a work are divided into two processes. Assuming that the user has not signed in, the first process involves the following steps (Figure 3):

- Step 1: From the EBI Database Website that is showing a specific dataset detail, the user clicks the 'sign in to ORCID' link, which triggers the sign-in process. In turn, the JavaScript THOR Client connects to the EBI ORCID Hub. If the user checks the 'Remember me on this computer' checkbox, the user will not need to go through the sign-in process again for a period of time.
- Steps 2 and 3: The EBI ORCID Hub redirects the user to ORCID for authentication and authorisation.
- Steps 4 and 5: Once authenticated and authorised, ORCID directs the user back to the EBI ORCID Hub with an authorisation code.
- Steps 6 and 7: The EBI ORCID Hub then exchanges the authorisation code for an access token.
- Step 8: The EBI ORCID Hub stores the access token in its database, together with the corresponding ORCID iD.
- Step 9: The EBI ORCID Hub then returns the user's information (such as name, and whether this dataset has already been claimed before) to the EBI Database Website, so that it can update the information displayed accordingly.

After authentication, the user can now claim his/her data by clicking the claiming link as shown in the process below, according to these steps (Figure 4):

- Step 1: The user clicks the claiming link, which instructs the JavaScript THOR Client to post the data entry to the EBI ORCID Hub.
- Step 2: The EBI ORCID Hub retrieves the corresponding access token from the database.
- Steps 3 and 4: The EBI ORCID Hub uses the access token to push the user's data to his/her ORCID record.
- Step 5: The EBI ORCID Hub indicates to the THOR Client whether the claiming was successful or not.

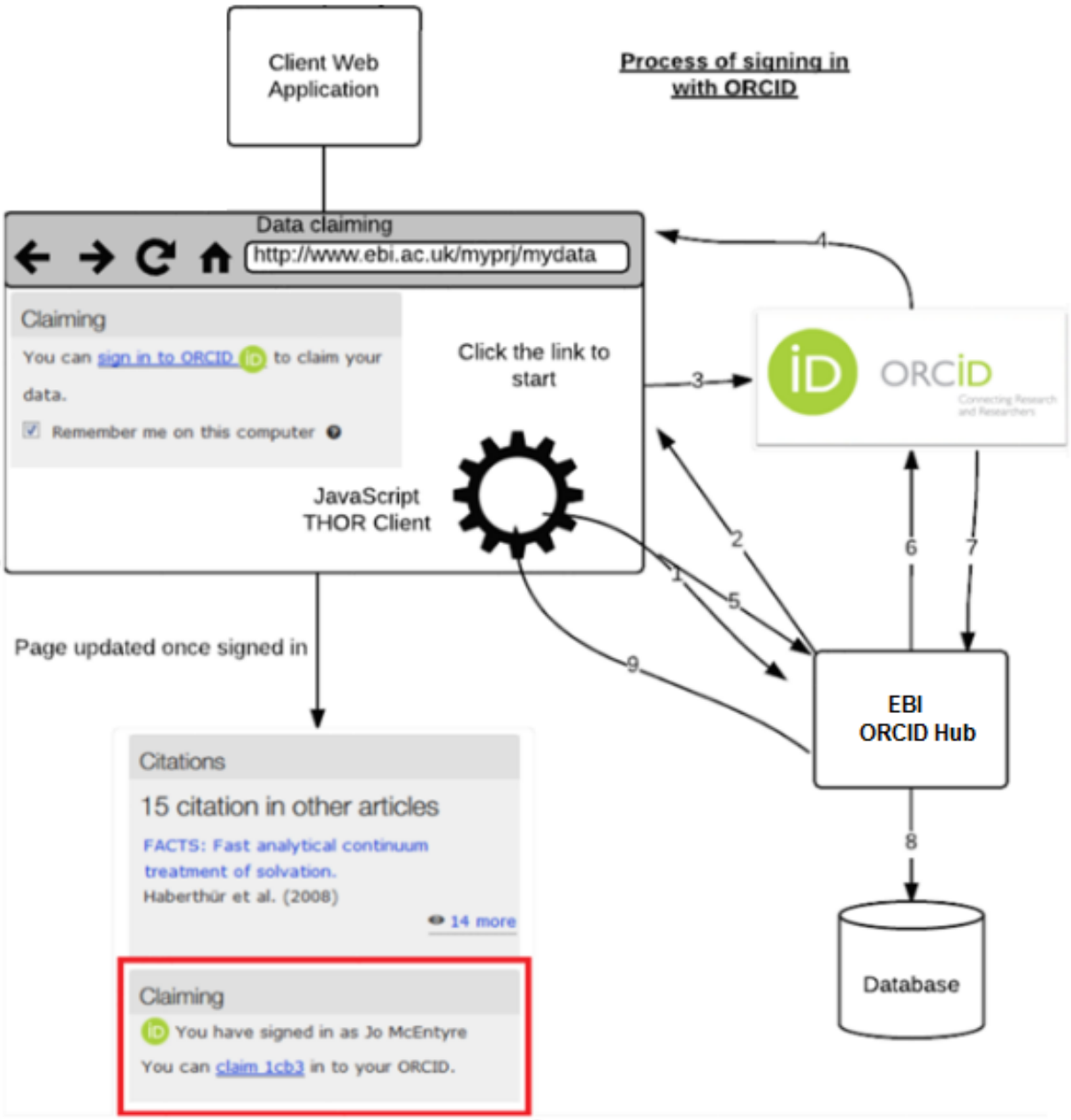

Figure 3: Sign in with EBI ORCID Hub

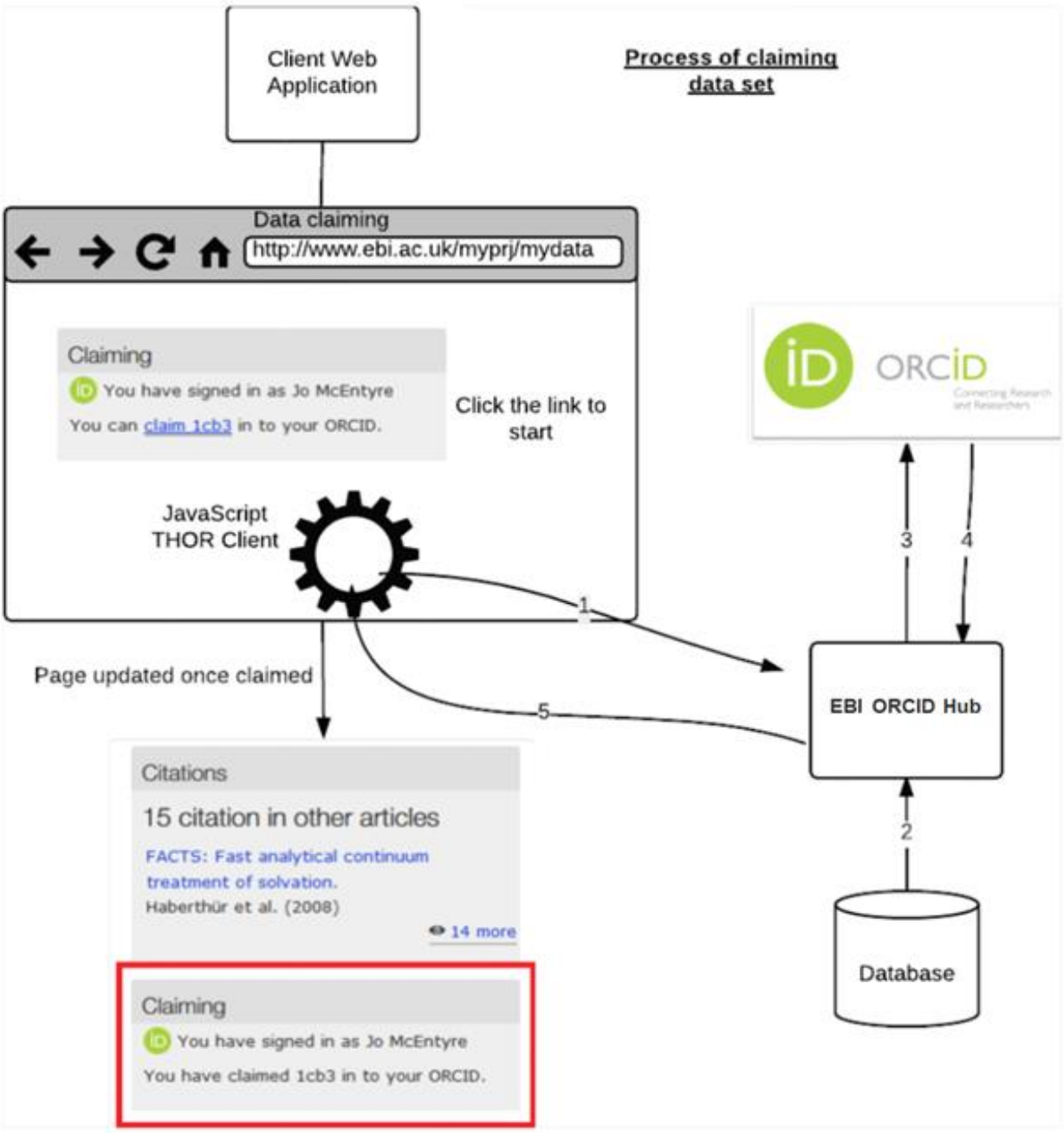

Figure 4: Dataset claiming with EBI ORCID Hub

The EBI ORCID Hub is built as a Java application running on a Tomcat Webserver, hosted at the EMBL-EBI. It uses Spring MVC for delivery of RESTful web services and Google OAuth Client library for authentication within ORCID. The features involved are:

- Define RESTful APIs to expose service functionalities
- Run client APIs over JavaScript and JQuery
- Transform the search results for data submission forms
- Manage the sign-in process through the OAuth 2.0 protocol
- Use the Hibernate framework to retrieve access tokens from the Oracle database
- Provide documentation

4.2 PANGAEA

Following the recent THOR-enabled integration of ORCID, PANGAEA now supports claiming of data publications to ORCID records. Strictly speaking, it is the contributor accounts that are claimed, and with this all associated contributor data publications. Hence, key to data publication claiming in PANGAEA is the association between a contributor account and ORCID iD. This association can be established in two ways: either by the contributor (self-claiming) or by PANGAEA (PANGAEA-claiming). PANGAEA is able to claim either manually or automatically.

PANGAEA data curators can claim manually by retrieving the contributor ORCID iD, and editing the contributor account accordingly. Such retrieval may occur via online search, or in direct communication with the contributor. However, in practice, PANGAEA curators rarely make use of this, as PANGAEA has also been operating the ORCID Resolver software algorithm in order to claim automatically. The ORCID Resolver algorithm was published open source (de Mello et al., 2016).

PANGAEA-claiming, either manually or automatically, relies on unvalidated ORCID iDs, as iDs are not obtained via the ORCID API following user authentication and PANGAEA authorisation at ORCID. PANGAEA-claiming is therefore not fail-safe, even though automated PANGAEA-claiming has worked reliably for PANGAEA over several years. The preferred approach to claiming is, instead, by contributors, i.e. self-claiming, since in contrast to PANGAEA-claiming it relies on validated ORCID iDs. This means that PANGAEA obtains a contributor's iD via the ORCID API after the contributor has authenticated their ORCID credentials and authorised PANGAEA to obtain information about them from ORCID (specifically their ORCID iD). Self-claiming requires that the user account and corresponding contributor account have been linked by a PANGAEA curator.

In order to self-claim, data publication contributors must create the association between their PANGAEA user account (which must have been previously created) and their ORCID iD by editing their PANGAEA user profile. Creating this association is functionality that PANGAEA has recently implemented as part of its ORCID Integration, and follows ORCID Best Practice (de Mello et al., 2016). Having established the association, PANGAEA then synchronises the corresponding contributor account with the ORCID iD. With this, contributors effectively self-claim all their data publications at PANGAEA. Figure 5 provides a graphical overview of (retrospective) data publication self-claiming at PANGAEA.

At PANGAEA, claiming is a process independent of data publication. Contributors can self-claim while data publications are being processed (i.e. pre-publication), or after datasets have been published (post-publication). The latter amounts to retrospective self-claiming and can occur at any point in time post-publication. As ORCID iDs are associated with user and contributor accounts, claiming is independent from individual data publications.

Establishing the association between the PANGAEA contributor account and the ORCID iD is the first step toward updating an ORCID record with data publications deposited at PANGAEA. Further steps are required, although these occur automatically without manual intervention from a contributor or curator.

Having obtained a contributor ORCID iD, PANGAEA annotates the data publication metadata accordingly. The annotation states explicitly the relationship between a data publication DOI and contributor ORCID iD. Together with other metadata, PANGAEA submits information about such relationships to DataCite. DataCite then shares this information with the wider network of PID infrastructures, specifically with ORCID. Figure 6 is a high-level overview for the metadata flow from PANGAEA to ORCID, via DataCite.

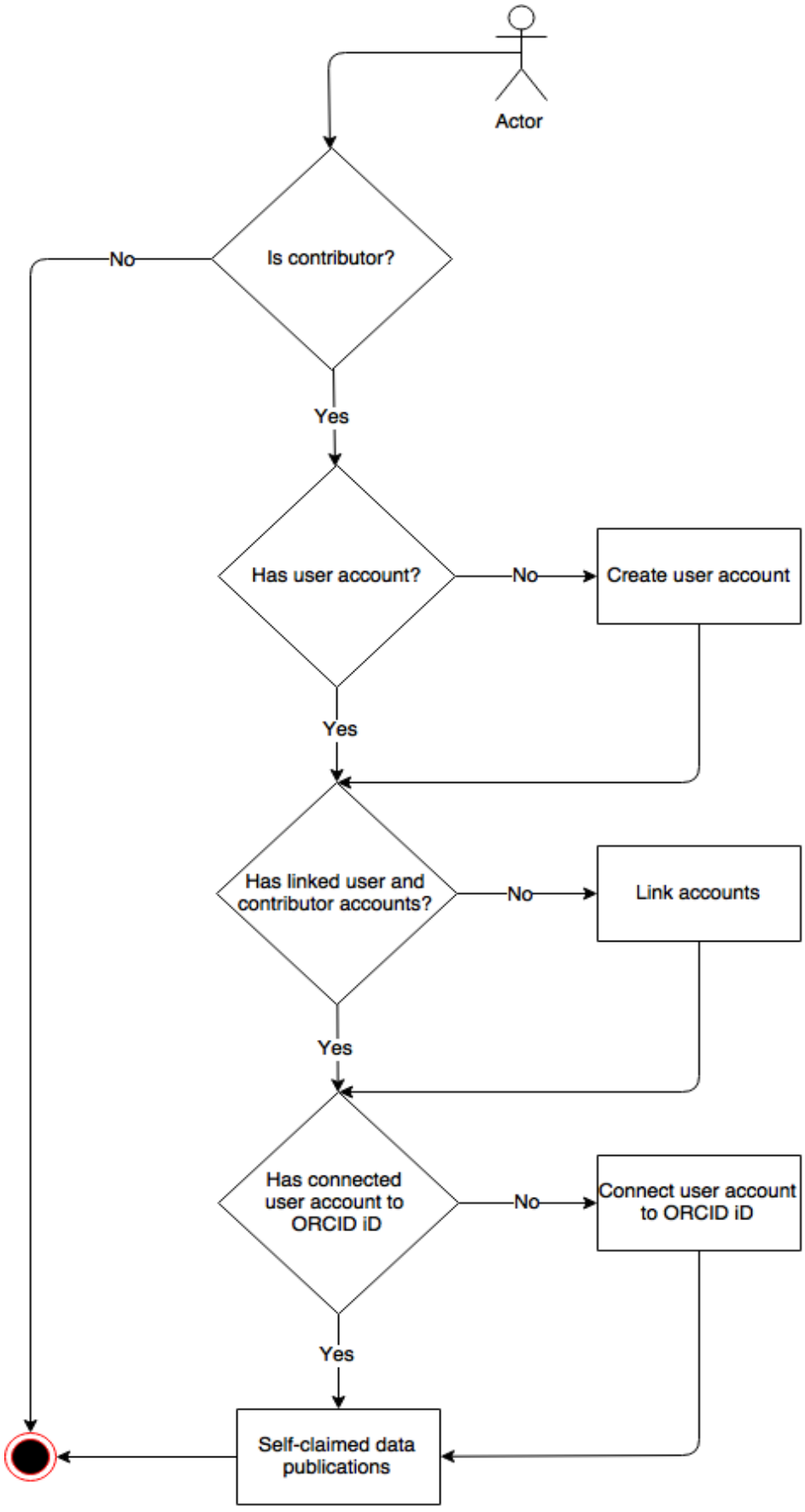

Figure 5: Activity diagram for (retrospective) data publication self-claiming at PANGAEA

Figure 6: Metadata flow from PANGAEA to ORCID via DataCite, in particular information about relationships between data publication DOIs and contributor ORCID iDs.

The choice for PANGAEA to push metadata to DataCite has both architectural and practical reasons. Architecturally, DataCite acts as a 'natural' hub for data centres such as PANGAEA, and shares information about PIDs among the wider network of PID infrastructures, including ORCID. With this multi-hub architecture (Burton et al., 2017), data centres only need to implement a communication interface with DataCite, and can thus delegate interoperability issues with the wider network of PID infrastructures to the hub. However, the choice to push metadata that is primarily relevant to ORCID from DataCite, rather than directly to ORCID, is also pragmatic, owing to the limitations that PANGAEA must work around because it is not an ORCID member. As a non-member, PANGAEA cannot push metadata directly to ORCID. Metadata is instead pushed to the DataCite Metadata Store.

The metadata is conformant with the DataCite Metadata Schema. Of particular interest here is how the relationship between a data publication DOI and a contributor ORCID iD is represented in metadata that PANGAEA pushes to DataCite. Listing 1 is an example XML snippet for the data publication with DOI 10.1594/PANGAEA.858171 and two contributors, Alice Lefebvre and Christian Winter.

As we can see in the example, PANGAEA data publication metadata includes the ORCID iD as a name identifier for any contributor with an association between contributor account and ORCID iD – so for any contribution for which the ORCID iD is known to PANGAEA. Unlike Christian Winter, Alice Lefebvre has associated her ORCID iD with her PANGAEA user account. Her ORCID iD is also associated with her contributor account because they had been linked by a PANGAEA curator. Consequently, metadata of data publications to which Alice Lefebvre has contributed explicitly represents the relationship between her ORCID iD and the data publication DOI. This explicit relationship in metadata represents the self-claim by the contributor, in this case by Alice Lefebvre, overriding the data publication. Together with the remaining metadata about the data publication, this relationship is pushed to DataCite and is then distributed within the wider network of PID infrastructures, specifically with ORCID. Based on this relationship provided by PANGAEA and shared by DataCite, ORCID eventually updates Alice Lefebvre's ORCID record with an additional Work for the PANGAEA data publication.

Metadata of *all* data publications authored by Alice Lefebvre (ORCID iD 0000-0002-9234-8279) will include the relationship between the DOI and the ORCID iD as shown in this example, not just this particular data publication. As noted earlier, contributors self-claim all their data publications by associating their ORCID iD with their PANGAEA user account. This is true also for retrospective self-claiming, which simply entails that, at any point post-publication, contributors sign in to PANGAEA with their user account credentials, and then associate the user account with their ORCID iD. A curator must then link the user account with the contributor account.

Listing 1: Example XML snippet showing the relationship between a data publication DOI and an ORCID iD in metadata that PANGAEA pushes to DataCite

```
<resource xsi:schemaLocation="http://datacite.org/schema/kernel-3 http://schema.datacite.org/meta/kernel-3/metadata.xsd" >
 <identifier identifierType="DOI" >10.1594/PANGAEA.858171</identifier>
  <creators>
   <creator>
    <creatorName>Lefebvre, Alice</creatorName>
    <nameIdentifier schemeURI="http://orcid.org" nameIdentifierScheme="ORCID" >0000-0002-9234-8279</nameIdentifier>
   </creator>
   <creator>
    <creatorName>Winter, Christian</creatorName>
   </creator>
  </creators>
 …
</resource>
```

4.3 CERN

The existing paper claiming service in INSPIRE offers a way for users to manage and review the accuracy of their automatically generated publications list. This feature is currently being moved to INSPIRE Labs as part of the overall update of INSPIRE. For the demonstration of data claiming functionality for HEP, this claiming service was extended to datasets on INSPIRE Labs as well.

Information on INSPIRE Labs is split into collections, with each collection appearing on a separate tab. For both the Literature and Data collections, authenticated users are presented with a list of the publications or data that the system believes to be theirs. Users can then indicate which papers are truly theirs by activating the toggles corresponding to each item. This claims these items to a user's author profile in INSPIRE and the flow is displayed on Figure 7.

Users have the option of associating their INSPIRE author profile with their ORCID iD and granting INSPIRE permission to push information to their ORCID profile. Once users have made this connection, works associated with the INSPIRE author profile will be regularly pushed to ORCID and appended to a user's ORCID record. Since the claiming service has previously been limited to papers, the works that were pushed to ORCID were likewise limited to papers. The inclusion of datasets as a claimable item will allow these to be pushed to ORCID in a similar way.

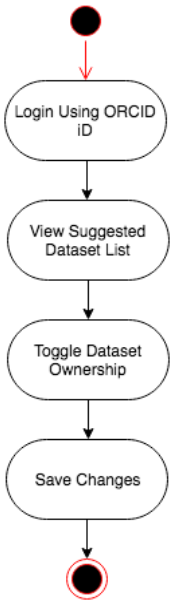

Figure 7: Activity diagram for claiming datasets in INSPIRE

4.4 DataCite

In order to address the DataCite data centre's needs, we updated the following DataCite services: Event Data, API and Search. Used in conjunction, these services synchronise aggregated information about claimed data back to data centres. We added extra functionality to the Event Data service to aggregate the user's claims, while new interfaces were added to the DataCite API and Search. There are two ways in which this information is presented.

1. Through the DataCite API, users can make calls to the services and obtain the claims information in machine readable format
2. Users can get information about claims in human-readable format via the DataCite search web frontend

4.4.1 DataCite API Implementation

Developers and system integrators can access the aggregated information about claims made by users in DataCite Search to the ORCID registry using the DataCite API. The DataCite API returns results in JSON, and follows the JSONAPI specification[14]. The desired information can be obtained by calling for `works` and filtering the results by `source-id`, which for this specific case is `datacite-search-link`. Additional filtering by data centre, member, publication year, and/or resource type (dataset, text, software, and so on) is possible using the query parameters `publisher_id`, `member_id`, `year`, `resource_type_id`, respectively.

For instance, users wishing to obtain information about datasets stored in Zenodo that have been claimed by authors in machine readable format can do so by calling the `works` API endpoint and filtering by `source-id=datacite-search-link`, `publisher-id=cern.zenodo` and `resource-type-id=dataset`:

> http://api.datacite.org/works?source-id=datacite-search-link&publisher-id=cern.zenodo&resource-type-id=dataset

At the time of writing, this query returns metadata for 199 datasets. Further documentation can be found in the DataCite API documentation[15].

4.4.2 DataCite Search Implementation

A user can access the aggregated information about claims made by users in DataCite Search to the ORCID registry using DataCite Search. For example, users wishing to obtain information in a human readable format about datasets stored in Zenodo that have been claimed by authors can do so by going to DataCite Search, clicking on the `Works` menu to see all DataCite DOIs, and then selecting the facets for `Resource Type: Dataset`, `Data Center: Zenodo` and `Source: DataCite (ORCID Search and Link)`:

> https://search.datacite.org/works?source-id=datacite-search-link&publisher-id=cern.zenodo&resource-type-id=dataset

Figure 8 shows example search results for the Zenodo data centre. This is essentially the same information as in the API example. Likewise, if someone is interested in looking at all claimed datasets by

[14] http://jsonapi.org/
[15] https://api.datacite.org/

a DataCite member organisation (for example, CERN in Figure 9) they could go to the Members menu, select the DataCite Member and filter by `Source: DataCite (ORCID Search and Link):`

https://search.datacite.org/members/cern?source-id=datacite-search-link

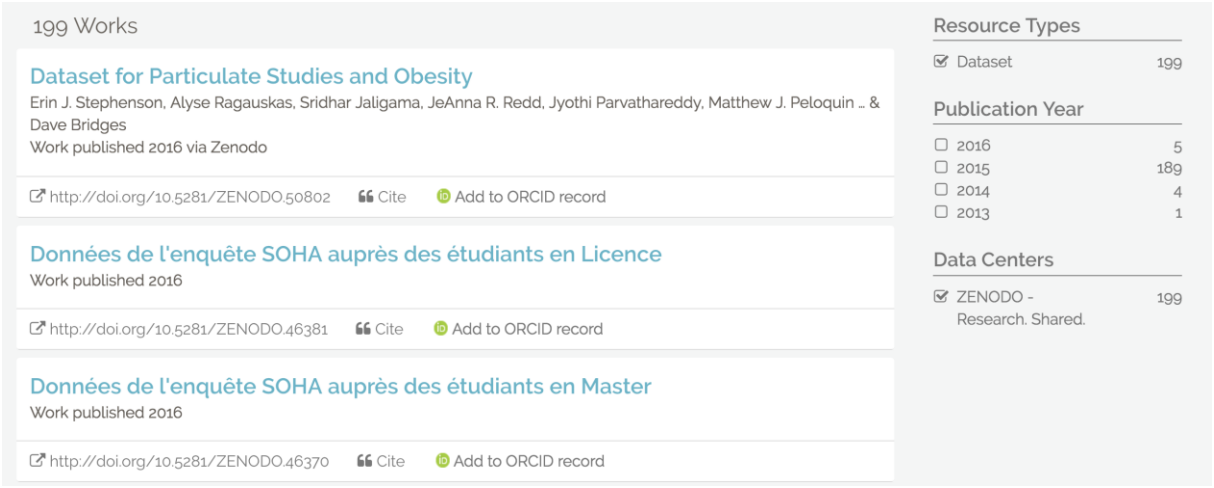

Figure 8: Search results for claimed datasets filtered by data centre (Zenodo)

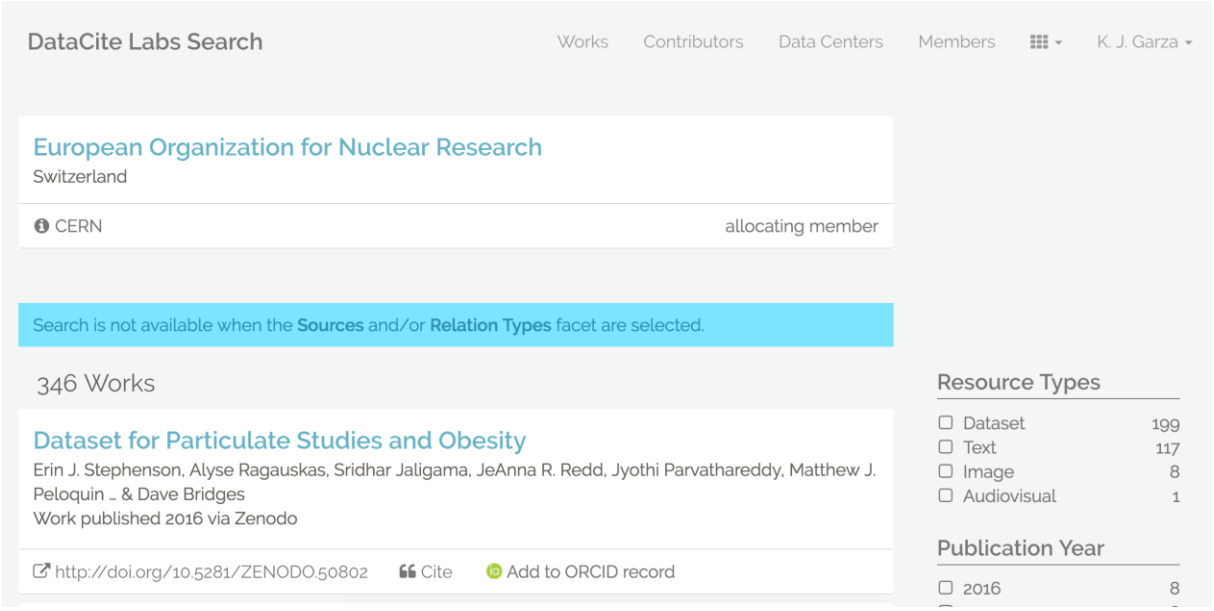

Figure 9: Search results for claimed datasets filtered by DataCite Member (CERN)

4.5 ORCID

Unlike PANGAEA, DataCite, and CERN, EBI uses a variety of domain-specific persistent identifier types (known as accession numbers). ORCID has worked with EMBL-EBI to review the way it deals with such external persistent identifier types. The existing situation was that they were described by an enumeration within the ORCID XML schema, meaning that adding new types was a breaking change for older clients, and only done with major new API releases. It was also a complex task that required significant developer effort.

As a result of this work, ORCID has performed major backend work to migrate these identifiers from the XML schema to a dynamic list in the soon to be finalised v2.0 API. This list is made available through the public API[16], and includes localised descriptions as well as the means to transform a non HTTP identifier into a resolvable URL. ORCID has also updated its open source contribution policy to include external identifier types, and there is a simple support desk procedure in place that members can use to request new identifier types within the ORCID registry.

Alongside this work, new types will be indexed automatically by our search tool so that they can be queried by interested data centres, enabling them to track references to their identifiers within the ORCID registry. Furthermore, these types will now be subdivided by the supported ORCID work relationship type, either 'self' or 'part-of', which will make querying chapters of books or journals, or subsets of data easier and more accurate.

5 Results

5.1 EMBL-EBI

EBI enhanced the existing service hub offered to its repository databases. It used to provide basic ORCID authentication features for database data submission forms: now it also allows users to claim works and datasets to their ORCID record. In addition, an ORCID search is now included in the hub to support databases in querying for ORCID users related to a particular work claimed.

Recently, EBI has been working with the databases to integrate with the EBI ORCID Hub in order to provide dataset claiming functionality to their users. The simplicity of the process has been emphasised: only a few updates on the part of the repository are required to use these new claiming features. Protein Data Bank (PDB) is the first EBI Database to make this integration; below we display screenshots from this resource to illustrate how the EBI ORCID Hub can be used. In addition to PDB, Metabolights is currently working on the integration, and several other EBI databases have committed to following suit, with the expectation that they will integrate with the service hub in the coming months, depending on agreements with their collaborators, governance structures, and development plans.

PDB 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).

[16] https://pub.orcid.org/v2.0_rc3/#!/Identifier_API/viewIdentifierTypes

Figure 10 shows the PDB website with integrated ORCID dataset claiming in the red box. Figures 11-13 highlight the integrated ORCID dataset claiming box in more detail. Figure 11 shows the claiming box before user sign-in, where the database can use the hub service to display a list of users who already have claimed that specific dataset.

Figure **12** shows what happens when the user clicks the sign-in link and is presented with a pop-up ORCID login form. Following sign-in, the service hub identifies the user, and checks with the ORCID registry if the user already has claimed the dataset being displayed, as shown in Figure 13. If the dataset has not yet been claimed, a link to claim the current dataset is shown. If the dataset is already claimed, a corresponding message is shown, as illustrated in Figure 14. Figure 15 shows the ORCID record of the user after the dataset is claimed, and information for dataset title, year, URL, and so on.

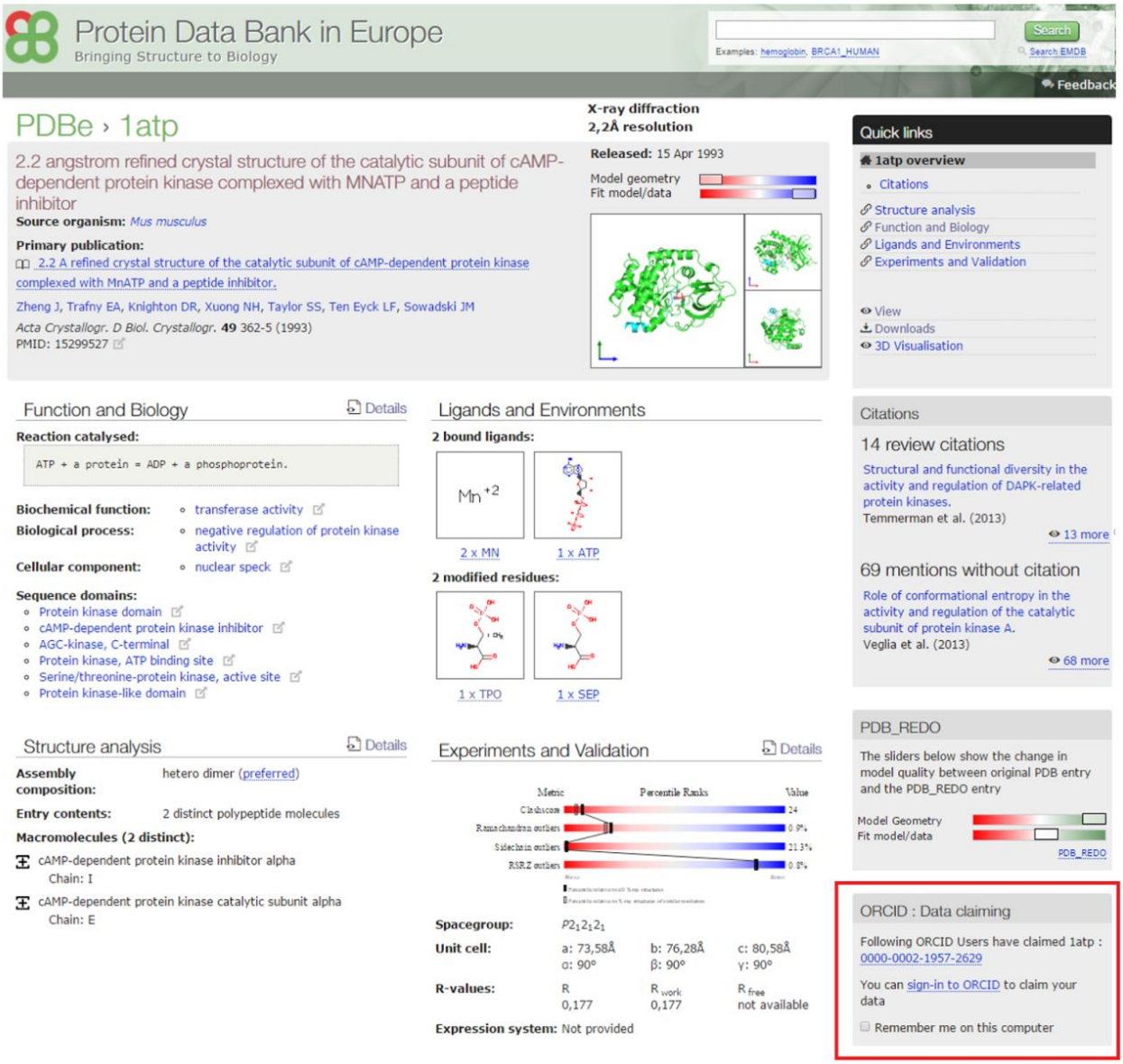

Figure 10: PDB website integrated with EBI ORCID Hub for dataset claiming

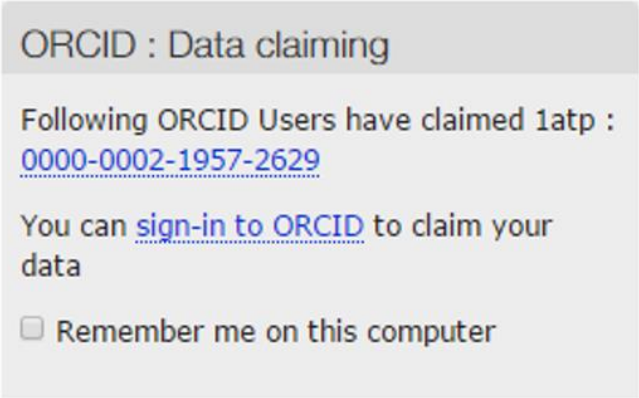

Figure 11: Dataset claiming box before user sign-in

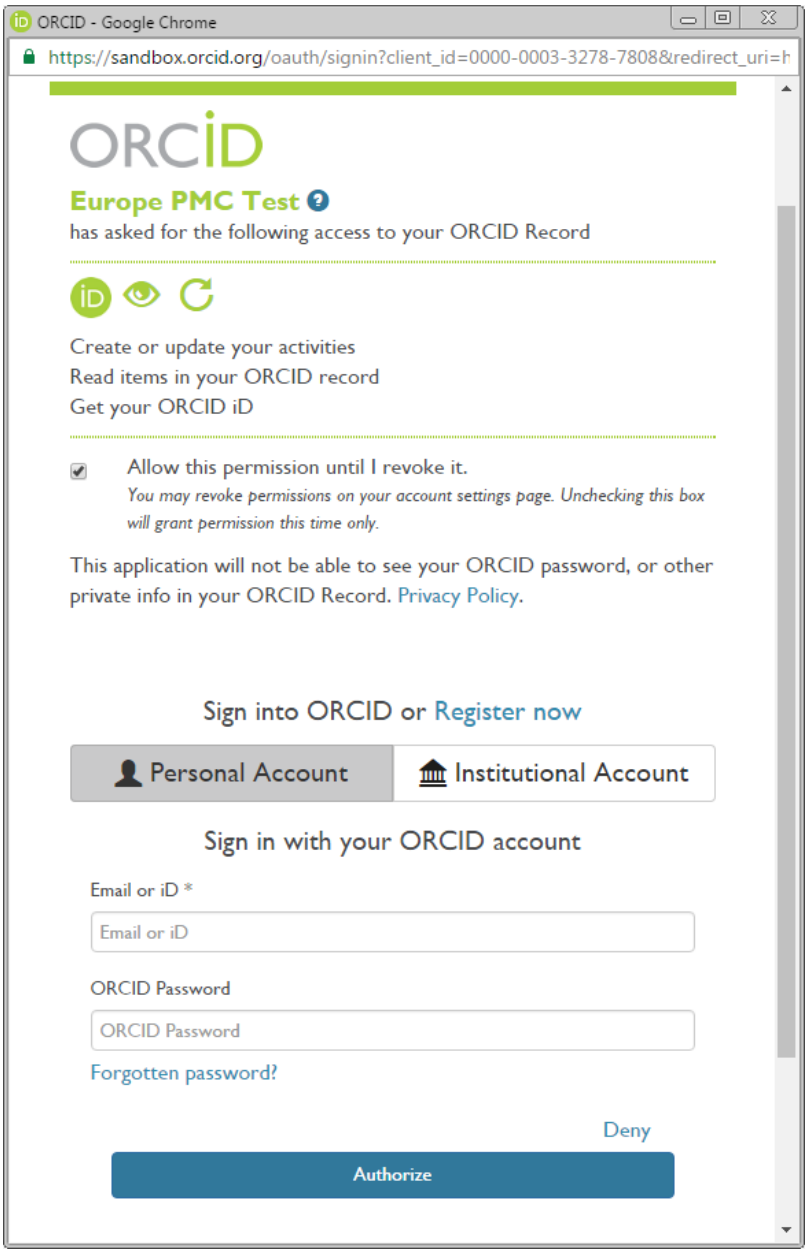

Figure 12: Redirection to ORCID for user authentication

Figure 13: Dataset claiming message for authenticated users

Figure 14: Dataset claiming message for claimed datasets

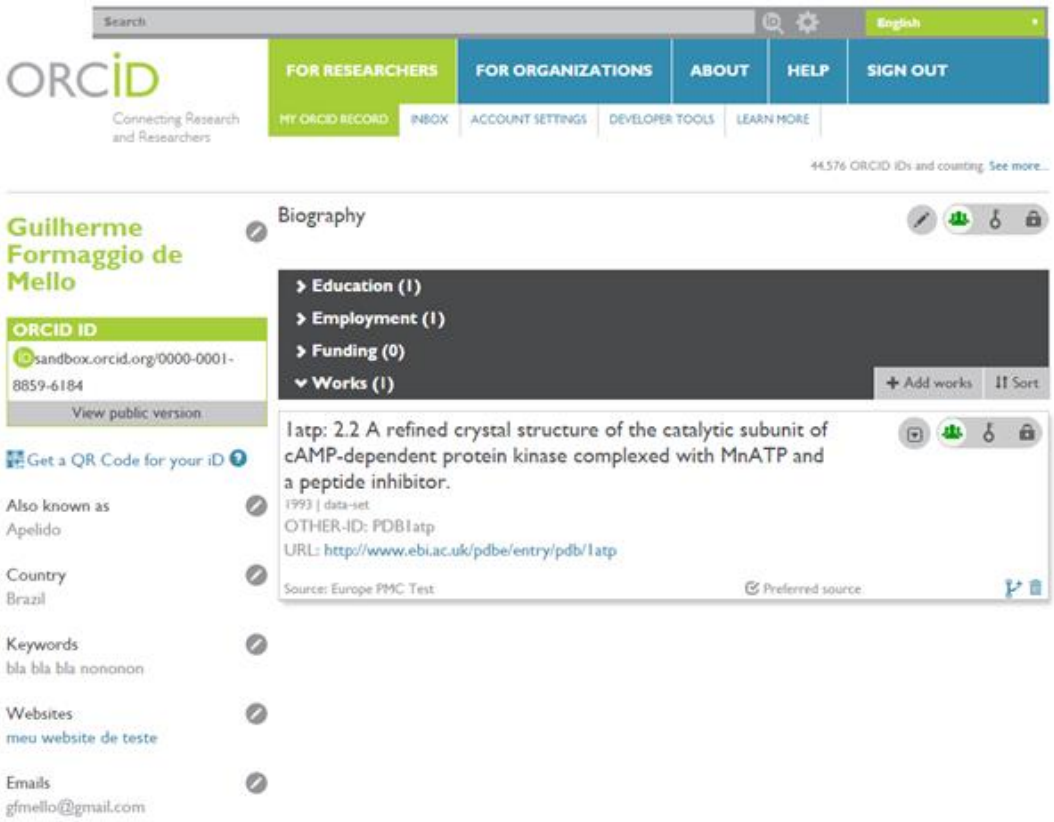

Figure 15: User's ORCID record with dataset information in the ORCID works section

5.2 PANGAEA

Building on the recently implemented and deployed ORCID integration, PANGAEA now supports (retrospective) self-claiming of data publications by contributors as works on their ORCID record. As discussed in Section 4.2, this is achieved by submitting metadata with information about relationships between data publication DOIs and contributor ORCID iDs to DataCite. Sharing such information with DataCite, and from there with the wider network of PID infrastructures, is the main result here.

Figure 16 shows an example PANGAEA data publication by Alice Lefebvre and Christian Winter (https://doi.org/10.1594/PANGAEA.858171). The landing page provides information about the publication. If available, the ORCID iD of contributors is shown by selecting contributor names in the citation. Figure 17 shows this feature more clearly.

Having created the association between Alice Lefebvre as a contributor to PANGAEA data publications and her ORCID iD, PANGAEA publishes this relationship between the DOI https://doi.org/10.1594/PANGAEA.858171 and ORCID iD http://orcid.org/0000-0002-9234-8279 to ORCID, via DataCite. As a result, ORCID displays metadata about the data publication on Alice's record (Figure 18). We have thus addressed the technical aspects of (retrospective) self-claiming of PANGAEA data publications to the ORCID record of contributors.

PANGAEA released its new website with ORCID integration and data publication claiming services on 30th September 2016. On 6th October 2016, PANGAEA informed its user base via email about the new website, specifically highlighting the new ORCID integration. This email reached out to 6,650 registered PANGAEA users. The new release was also advertised on Twitter.

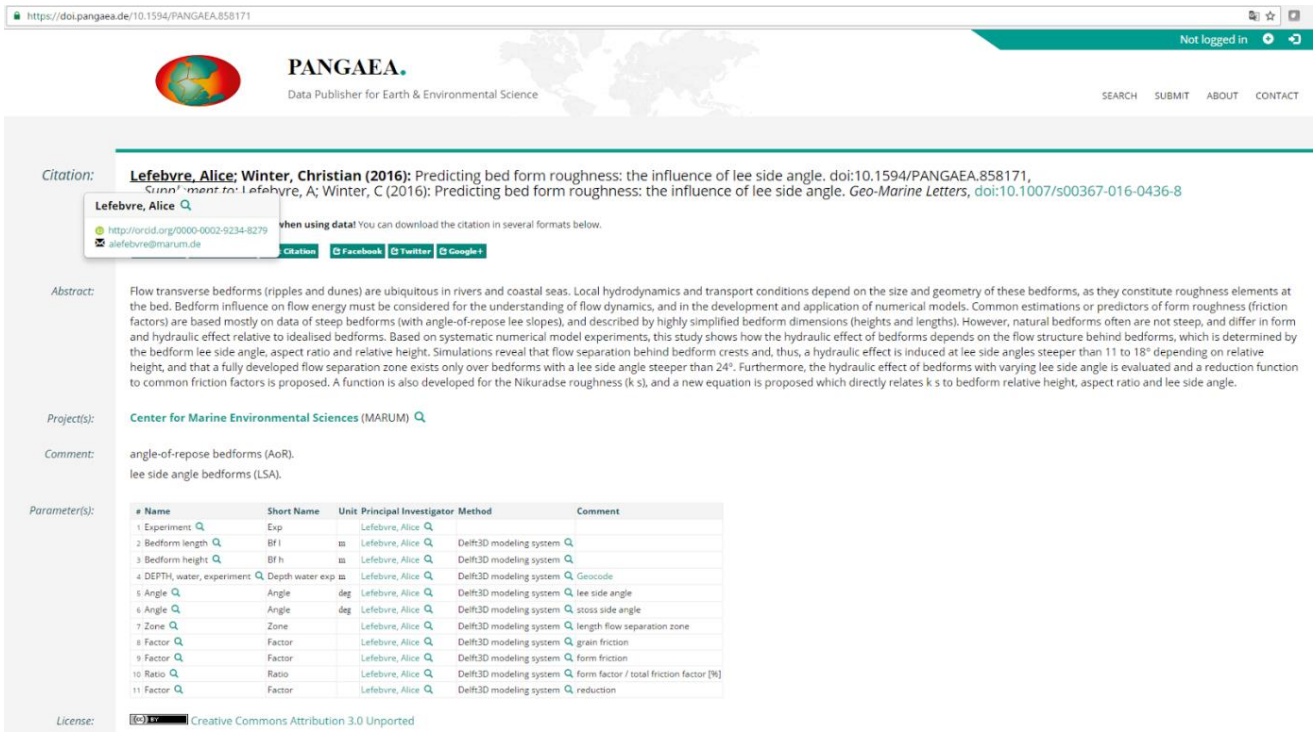

Figure 16: Data publication by Alice Lefebvre and Christian Winter with DOI https://doi.org/10.1594/PANGAEA.858171. The ORCID iD of contributors is shown by selecting contributor names in the citation.

Figure 17: Alice Lefebvre's ORCID iD shown by selecting her name in a data publication citation

It is too soon to monitor trends in the utilisation of these new services by PANGAEA users. Currently, 6.6% of PANGAEA contributor accounts have an associated ORCID iD; 16.5% of data publications have at least one contributor with an associated ORCID iD; and 7.1% of data publications have all contributors with associated ORCID iD. We will monitor these and other related figures over time and try to obtain trends.

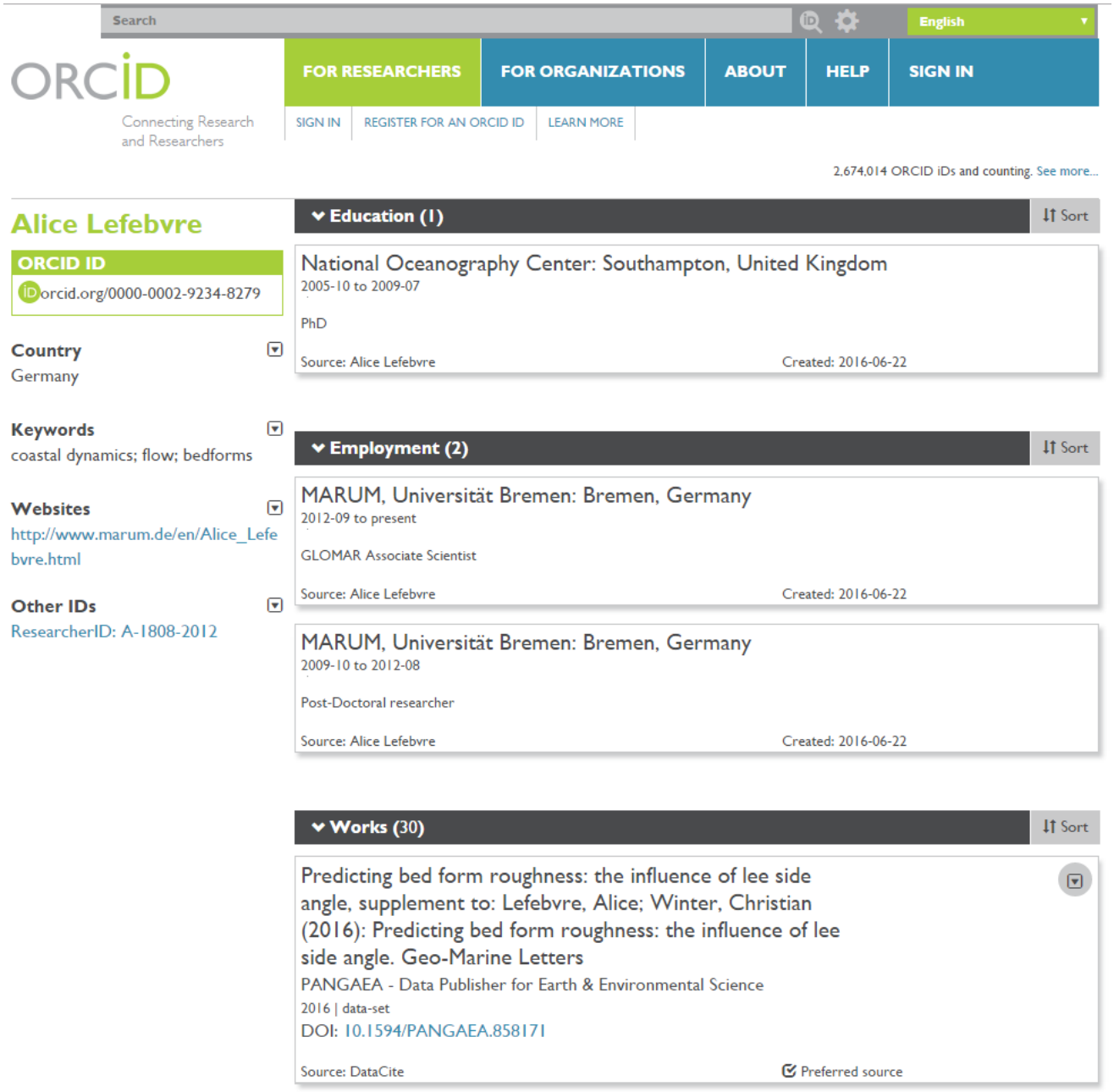

Figure 18 Alice Lefebvre's ORCID record showing her PANGAEA data publication with DOI https://doi.org/10.1594/PANGAEA.858171 as an ORCID work

5.4 CERN

Extending an existing service to include an additional item type may be conceptually straightforward, but it is not without challenges in INSPIRE's current environment. First and foremost, INSPIRE is in the middle of a lengthy process of being upgraded. Invenio, the underlying platform for INSPIRE, is concurrently being upgraded to version 3.0, which entails an entire re-envisioning of its architecture into a more modular form. The new INSPIRE is being released in stages, with upgraded features being released on INSPIRE Labs.

As stated previously, INSPIRE Labs is divided into collections, such as Literature and Data. Literature displays search results for papers from within INSPIRE, while the Data tab displays a view of search results from HEPData. In keeping with the nature of the previously existing relationship between INSPIRE and HEPData, no information is actually ingested from HEPData; it is only displayed in the INSPIRE context.

In both the Literature and Data collections, users can search for records that meet certain criteria, such as those authored by a particular person, including themselves.

An authenticated user will be presented with an additional option to enable Claim Mode via a toggle. When enabled, Claim Mode filters the current result set with the authenticated user's name. Claim Mode then allows the user to select records for items that they authored via toggles to the right of each item in the search results list.

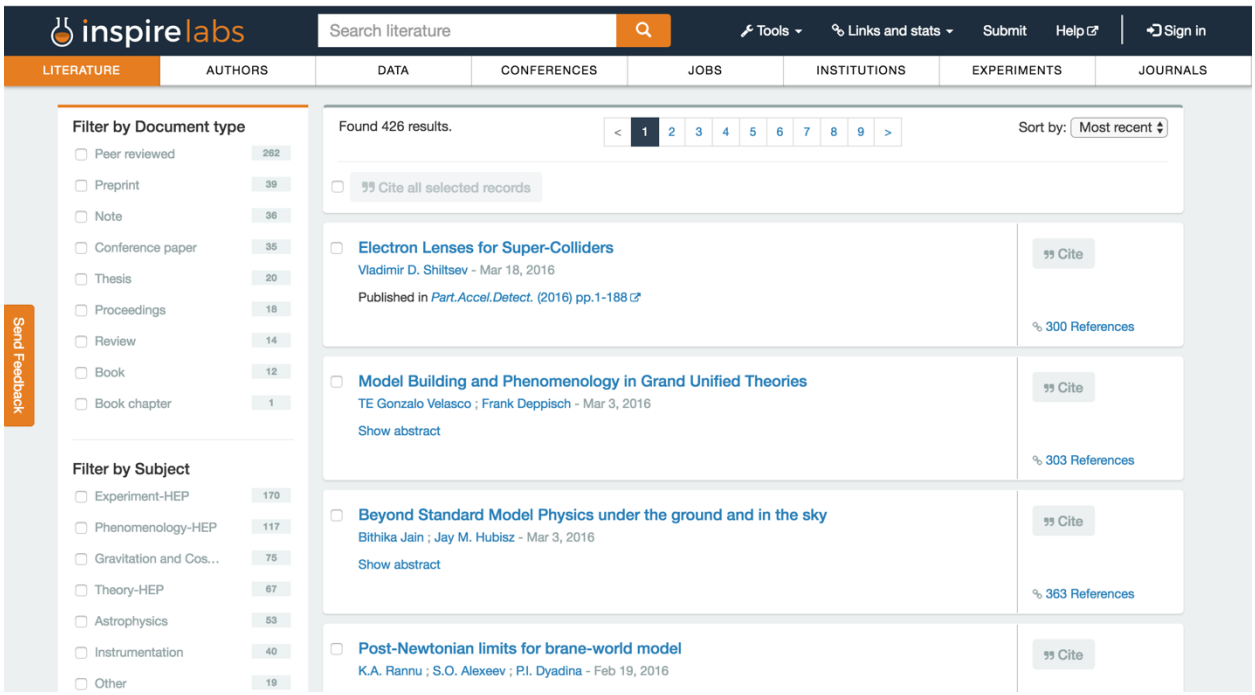

Figure 19: Search results on INSPIRE Labs

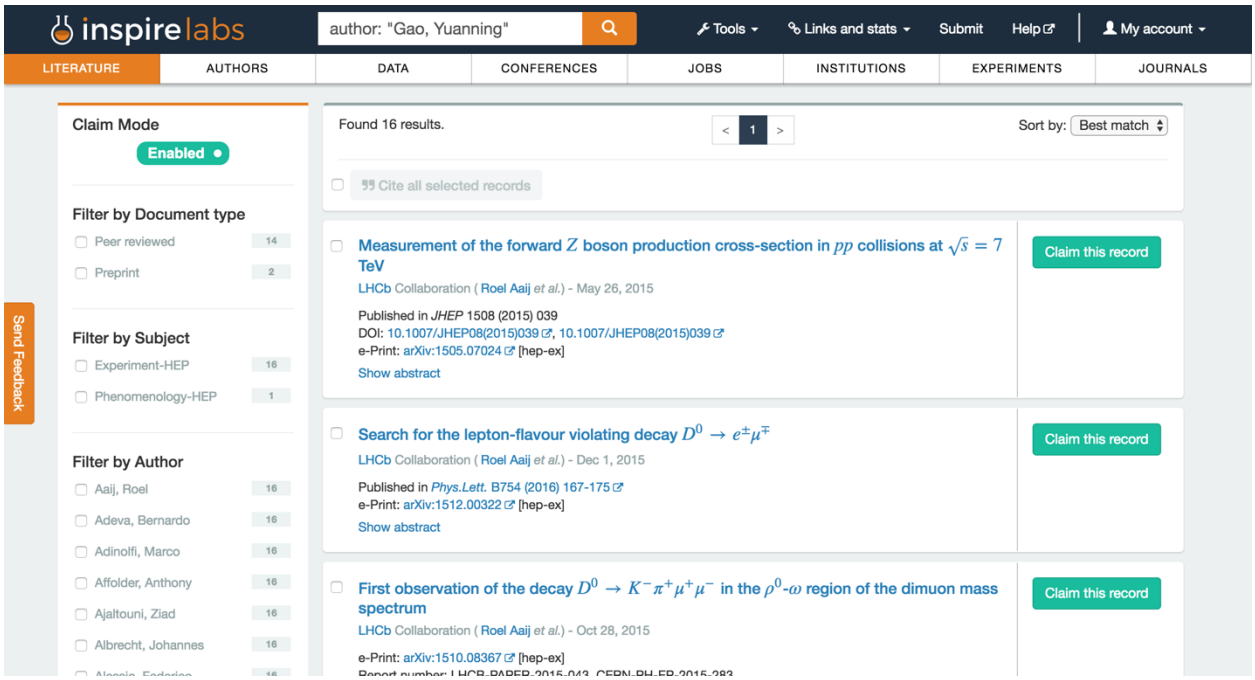

Figure 20 : Claim Mode enabled for an authenticated user on INSPIRE Labs

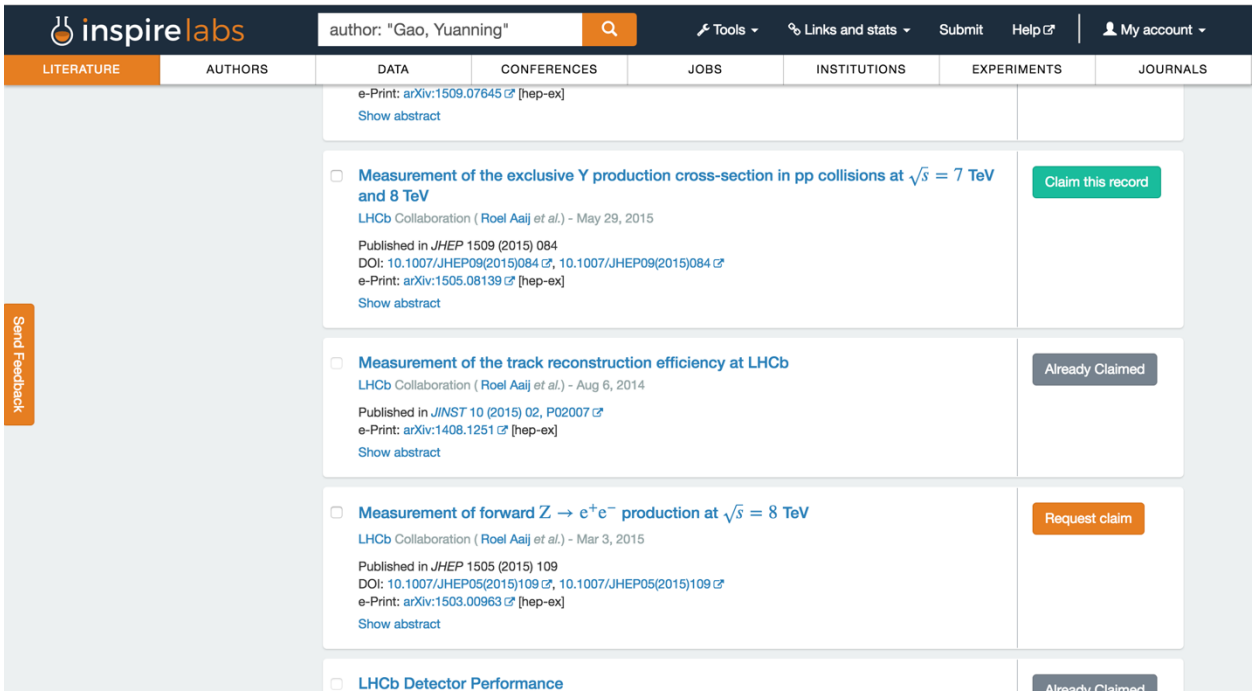

Figure 21: Examples of possible claim status indicators while in Claim Mode on INSPIRE Labs

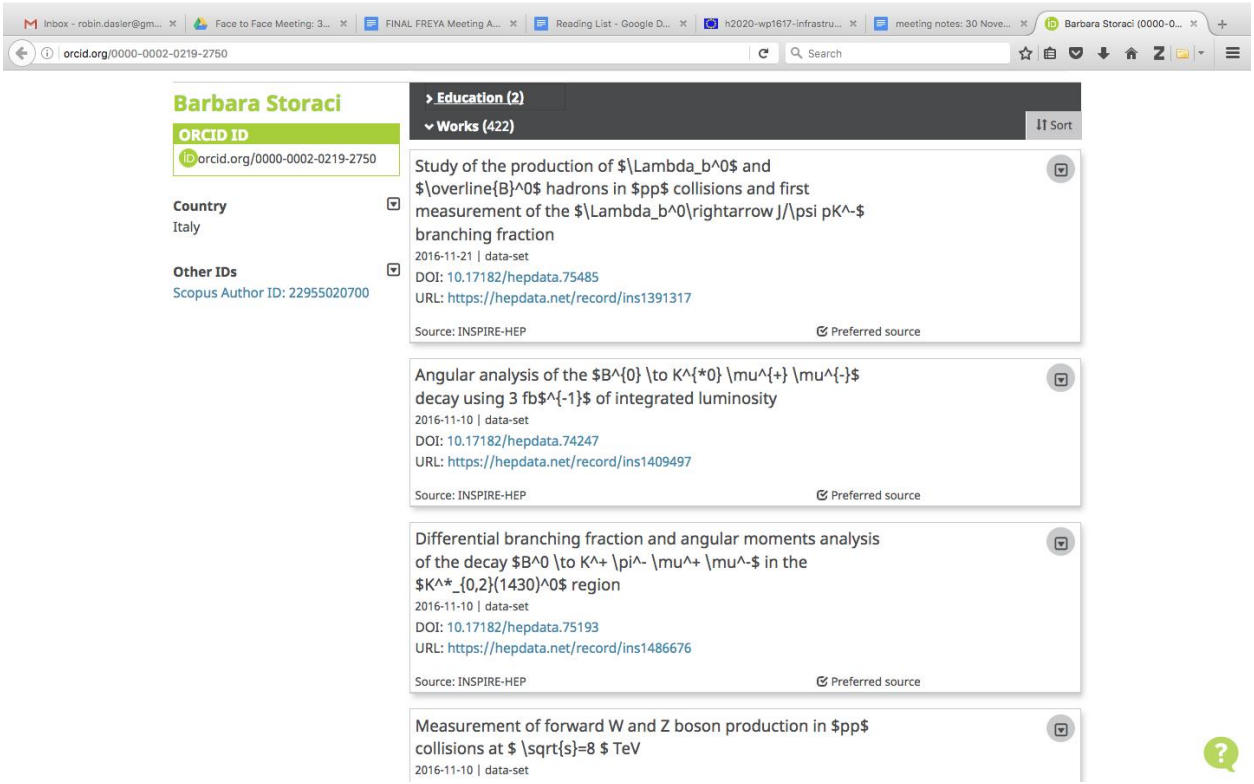

Figure 22: Datasets in a user's ORCID record that were pushed from INSPIRE Labs after claiming

Items selected by the user are then claimed to their INSPIRE profile, which will in turn append these items to the user's ORCID record as part of INSPIRE's existing ORCID push operations. Because the new version of INSPIRE will offer ORCID authentication as the only login option, we are assured that those claiming items to their INSPIRE profiles will have an associated ORCID iD.

5.5 DataCite

From a technical point of view, the changes in the DataCite API and Search front-end provide a number of important new features. They provide a clear presentation of claims when listing all works by data centre or DataCite member as well as providing faceted functionality by publication year and resource type. These claims are made by users in DataCite Search (Figure 23) and are sent to the ORCID registry. This new functionality allows data centres to import these claims and to add them to their own systems, enriching the information about their datasets with ORCID IDs.

DataCite faced some technical challenges with the solutions that we explored. An idea we explored but postponed was the implementation of a mechanism that notifies data centres when an author claims a dataset. This notification mechanism would seamlessly share information about the ORCID record that was updated via DataCite's services. However, the potential technical solutions that we considered (for example, email notifications and webhooks) required significant resources from data centres. For example, email notifications are restricted by the availability of contacts from data centres and defining an overall strategy on notifications. On the other hand, webhooks uptake is limited by a data centre’s level of sophistication to deal with this technical solution.

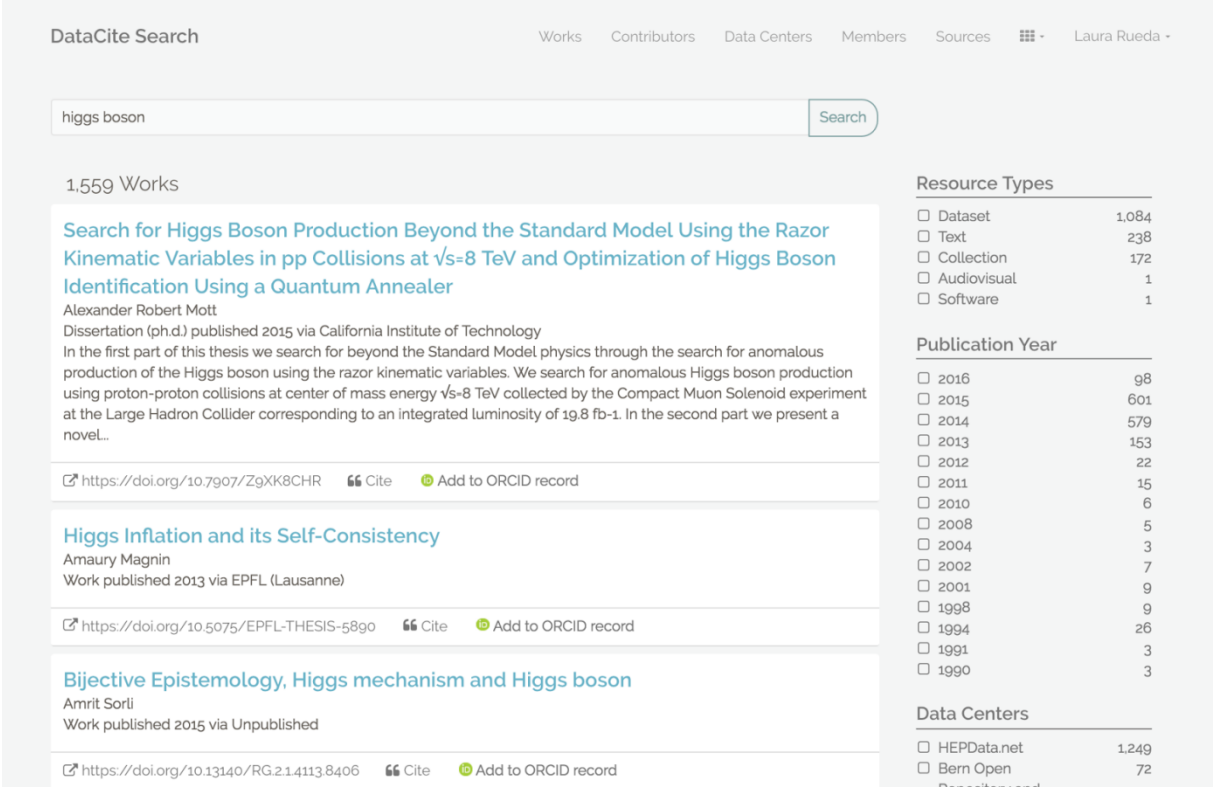

Figure 23: DataCite search and facets

The implementation we present here requires few resources from data centres, but still poses some limitations. These restrictions are related to additional functionality of the existing DataCite services. These limitations are: a) faceting by `source` or `relation_type` is slower as it hits our relational database directly; and b) the user cannot perform queries (by keywords) when faceting by `source` or `relation_type`, also because these queries are executed directly on our relational database rather than on the Search index. DataCite will address these limitations in future releases of the DataCite API and Search services.

5.6 ORCID

ORCID's move to a more dynamic list of identifier types has enabled EBI to better represent their work types within the ORCID registry. In the first instance this is with protein databank identifiers (PDB), but it is anticipated that new types will be added as more EBI services integrate ORCID. Other services are already requesting new identifier type support, and, at the time of writing, KoreaMed and lens.org identifiers are both in the deployment pipeline.

The improvements to search indexing mean that it is now easier for these services to understand who is claiming which research objects within their repositories by querying for their own identifier types directly. The improved search will also help to propagate the new cross-links that emerge from an expanded set of identifier types, as more objects are represented by not only general purpose PIDs such as DOIs or Handles, but also their institutional/domain specific identifiers.

Wider implications of the work include reducing the barrier to upgrading ORCID API versions by removing the dependency on the XML schema for identifier type validation, reducing the development effort required to add new identifier types, and making ORCID more responsive to community identifier requirements.

6 Challenges and Lessons Learned

There is a shared understanding among the institutions that have contributed to this report that the remaining challenges to dataset claiming are largely of a social, rather than a technical nature. EMBL-EBI, PANGAEA, CERN, and DataCite have addressed the technical aspects of dataset claiming in their respective workflows. While there are almost certainly further technical details to be addressed and ironed out, the more challenging barriers are in the adoption of ORCID by the respective user bases as well as in researchers actually claiming datasets.

6.1 PANGAEA

Each institution – as well as the community of data centres as a whole – arguably has a number of potential ways in which to motivate researchers to create ORCID iDs and perform claiming. PANGAEA strongly encourages its users to connect accounts to ORCID iDs. The encouragement takes the form of a message visible to users in profile editing. Furthermore, as PANGAEA curators are in personal contact with the contributors who submit data to PANGAEA, they could personally inform the submitter about ORCID, and the potential of connecting a user account with an ORCID iD. As PANGAEA has the email address of the submitter, such information could also be provided automatically.

A more stringent approach would be to require the inclusion of the ORCID iD during sign up. This would meet the THOR recommendation that 'user accounts should include an ORCID iD' as well as reflect a notable trend among article publishers that have started to require the ORCID iD of the corresponding author.[17] Making the ORCID iD a required field to sign up for a PANGAEA user account would arguably facilitate claiming; however, in practice such a requirement may be too strict, as not all users have an ORCID iD, or worse are not willing to obtain one at the time of signing up for a PANGAEA user account.

For contributors other than the submitting author, one possibility could be to encourage the submitter (for example, in direct communication with curators) to inform co-authors about ORCID and the possibility of connecting their ORCID iD and PANGAEA user accounts. If PANGAEA has obtained co-author emails, another possibility is to inform and invite co-authors directly. This is an approach Elsevier has adopted recently for (some of) its journals.

PANGAEA currently does not require the submitting author to provide email addresses for all authors of data publications. Making email information mandatory could help to automate invitations to all contributors of data submitted to PANGAEA – both in pre-publication as well as potentially in post-publication. The most extreme scenario could be to place the data publication on hold until all co-authors have created a user account and connected the account to their ORCID iD. With this approach, all data publications would be effectively self-claimed by all authors at the time of data submission. However, such a strict requirement is, in practice, hardly viable and could be counterproductive, resulting in submitted data failing to be published because a co-author is unwilling to self-claim the data publication using ORCID.

6.2 CERN

CERN faces similar social challenges of encouraging user uptake of ORCID iDs. As a means of encouraging ORCID use, the new version of INSPIRE will offer ORCID authentication as the sole means of authentication

[17] http://orcid.org/content/requiring-orcid-publication-workflows-open-letter

for those functions that require a login; in the current version of INSPIRE, it is possible to use an arXiv login as an additional option. The vast majority of functionality on INSPIRE is, however, available without logging in: only claiming and suggestions/corrections require a login. As previously stated, author profiles are automatically generated from harvested papers, so it is technically possible for a HEP author to have a profile without any authentication-requiring maintenance.

The remaining challenges for data in INSPIRE are largely philosophical. INSPIRE is an aggregator, not a repository. Its primary function is to harvest information that exists elsewhere and provide that information in one place to a specific disciplinary community of users. This means that the harvested information by and large does not reside with us. This general philosophy also extends to INSPIRE's relationship with HEPData. Responsibility for the datasets resides in HEPData, but the two systems share metadata back and forth for mutual benefit. The inclusion of directly-uploaded datasets in INSPIRE is a historical anomaly born of last resort that does not completely align with current philosophy. The solution presented herein is one step toward re-aligning practice and philosophy on our path to an improved INSPIRE product.

6.3 DataCite

The social challenge for DataCite comes at a different level of granularity than the challenges from CERN, EBI, and PANGAEA. The challenge for DataCite is helping to integrate DataCite's services that could potentially motivate authors and researchers to adopt the ORCID iD. For example, in contrast to data centres that have implemented ORCID integration, the login to DataCite services requires an ORCID iD to be associated with the account. The challenge for DataCite is to achieve service integration with all of its data centres.

6.4 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 requires the participation of many databases. Therefore, as the host for many of the key data resources in Europe, the EMBL-EBI needed to develop a platform of services that can be integrated into the many EBI-hosted databases. This was accomplished by the creation of the EBI ORCID Hub to enable ORCID dataset claiming in each of its systems. While this was relatively easily 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.

6.5 Contributor Roles

Systems could allow claiming under different roles. Issues that arise in this context are: the use of contributor roles; uptake of defined roles for attribution within our respective communities; more fundamentally, whether or not contributor roles are even desired by our communities; and whether advocating for their use would actually be best practice in a disciplinary context. While the discussion is still ongoing, the current consensus is that our respective disciplinary information systems are not prepared to take on this risk (if indeed it is a risk) in our development at this time. That being said, none of the solutions outlined in this document by THOR's disciplinary partners necessarily preclude the use of contributor roles. These roles are simply not part of the metadata that we are pushing at this time.

6.6 Synchronising Claimed Data

On the question of *implementing a mechanism to synchronise claimed data back to multiple contributing data centres,* relevant issues are:

- Who should be responsible for executing this synchronisation
- Whether the best way to propagate this information is by 'pushing it' to the repositories through a notification mechanism

ORCID's existing notification functionality is a premium membership feature that enables interested parties to register to receive webhook notifications whenever a specific ORCID record is modified. This is consistent with ORCID's researcher centric view of the scholarly communication landscape, and its sustainability model of providing value added features to paying members. Similarly, DataCite provides research object centric value added services to its own members. DataCite's membership encompasses an extensive network of relationships within the data centre community, whereas ORCID does not. As a result, DataCite is better placed to provide them with services.

In many instances push notifications are unsuitable for propagating dataset claims back to the contributing data centres; other means, such as API calls or even emailed CSV, are sometimes more appropriate mechanisms of exchanging data. This is a result of a long tail of legacy databases that cannot take advantage of webhooks.

With this in mind, the team decided that DataCite would implement these research object orientated services for discovering claims, which are described earlier in this report. ORCID supplemented this with improvements to the way in which these links can be discovered within ORCID, providing improved 'pull' functionality for those who require it.

7 Conclusion

The THOR partners have developed services that allow users of their respective systems to claim datasets to their ORCID profiles retrospectively. This work was largely enabled by – and in most cases built on top of – previously completed THOR-facilitated work to integrate ORCID authentication into our systems.

Although each of the disciplinary partners had unique concerns driven by differently constructed service ecosystems, all were able to successfully implement claiming services that would suit their individual requirements and environments. Furthermore, the infrastructural improvements developed by DataCite and ORCID have provided, and will continue to provide, a stronger backbone for future development of interoperable services, enabling other data centres to implement similar services appropriate for their unique environments.

The work reported in this document shows true progress in implementing claiming processes in multiple organisations and disciplines. More organisations can build on our experience and get started right now. There are, of course, some significant technical challenges to improve workflows, such as synchronising information about claims across service providers, or scaling to handle very large numbers of claims. Now that claiming processes are being implemented, however, additional attention must be devoted to the human and social factors so that automated, accurate associations between people and their research outputs become the new normal.

8 References

Burton, A., Koers, K., Manghi, P., Stocker, M., Fenner, M., La Bruzzo, S., Aryani, A., Diepenbroek, M., Schindler, U. (2017). The Scholix Framework for Interoperability in Data-Literature Information Exchange. *D-Lib Magazine*, Volume 23, Number 1/2. https://doi.org/10.1045/january2017-burton

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

Europe PMC Consortium. Europe PMC: a full-text literature database for the life sciences and platform for innovation. Nucleic Acids Res. 2015 Jan;43(Database issue) D1042-8. https://doi.org/10.1093/nar/gku1061. PMID: 25378340; PMCID: PMC4383902.

Appendix A: Project Summary

The **THOR** project establishes a sustainable international e-infrastructure for persistent identifiers that enables long-term access to critical information about the life cycle of research projects. It enables seamless integration between articles, data, and researcher information, creating a wealth of open resources. This will result in 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[18]).

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

[18] https://www.force11.org/group/joint-declaration-data-citation-principles-final

Appendix B: Terminology

Additional terms are defined below:

Term	Definition
API	Application programming interface
arXiv	Open access e-print archive (Physics, Mathematics, Computer Science, Quantitative Biology, Quantitative Finance and Statistics)
CERN	CERN, the European Organization for Nuclear Research, is one of the world's largest centres for scientific research. http://home.cern/
Crossref	Digital Object Identifier Registration Agency for scholarly publishing
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
EC	European Commission
EMBL-EBI	European Bioinformatics Institute , part of the European Molecular Biology Laboratory
Europe PMC	Repository providing access to worldwide life sciences articles, books, patents and clinical guidelines
HEP	High Energy Physics
ID	Identifier
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting
ODIN	ORCID and DataCite Interoperability Network
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.
ORCID iD	Persistent digital identifier that distinguishes individual researchers and, through integration in key research workflows such as manuscript and grant submission, supports automated linkages between individuals and professional activities.
PANGAEA	Data Publisher for Earth & Environmental Science
PDB	Protein Data Bank
PID	Persistent Identifier
PMC	PubMed Central