



## Demonstration of Services to Integrate ORCID IDs into Data Records and Database Systems

### Document Information

- Date:** 28/07/2016
- Authors:** Guilherme de Mello (EMBL-EBI), Florian Graef (EMBL-EBI), Markus Stocker (PANGAEA), Uwe Schindler (PANGAEA), Robin Dasler (CERN), Johanna McEntyre (EMBL-EBI), Sünje Dallmeier-Tiessen (CERN)
- Reviewers:** Adam Farquhar (British Library), Rachael Kotarski (British Library)
- Abstract:** This report summarises progress on integrating ORCID iDs into production services in major databases at three different organisations and disciplines, namely EMBL-EBI for life sciences, PANGAEA for earth sciences, and CERN for high-energy physics. ORCID integration is enabling these services and databases to automatically link deposited datasets with the unique and persistent identifier of contributors, thus facilitating unambiguous credit for the production of datasets. We discuss the requirements, challenges and lessons learned.
- DOI** 10.5281/zenodo.58971

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 2016.

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



## Contents

Executive Summary .....	1
1. Introduction .....	1
2. The Institutions and their Services .....	3
2.1 ORCID: Open Researcher and Contributor ID .....	3
2.2 EMBL-EBI: The European Bioinformatics Institute .....	3
2.3 PANGAEA .....	5
2.4 CERN .....	5
3. Requirements for Integrating ORCID IDs into Data Records .....	7
3.1 General Implementation of ORCID Authentication .....	7
3.2 Institution-Specific Requirements .....	8
3.2.1 EMBL-EBI .....	8
3.2.2 PANGAEA .....	8
3.2.3 CERN .....	9
4. Work Completed .....	10
4.1 EMBL-EBI .....	10
4.2 PANGAEA .....	12
4.3 CERN .....	14
5. Results .....	16
5.1 EMBL-EBI .....	16
5.2 PANGAEA .....	19
5.3 CERN .....	23
6. Challenges and Lessons Learned .....	26
6.1 Open Development and Code Sharing .....	26
6.2 Considerations on Authorship and Submission Workflows .....	27
7. Conclusion .....	28
Appendix A: Terminology .....	29
Appendix B: THOR Project Summary .....	30



## Executive Summary

Researchers, their contributions, and their scientific artefacts need to be unequivocally connected. They face the ongoing challenge of distinguishing their research activities from those of others with similar names. They need to be able to easily link unique and persistent identifiers with their research artefacts: publications, datasets, equipment, articles, media stories, citations, experiments, patents, notebooks and more.

To facilitate connections between researchers and data, THOR is developing ways to support the inclusion of ORCID IDs in databases and services. This document reports on successful integration of ORCID in databases and services of three partners serving three distinct disciplines, namely PANGAEA (earth sciences), EMBL-EBI (life sciences), and CERN (high-energy physics). Each partner has enabled the preferred method of ORCID ID inclusion during data submission, with data submitters authenticating via ORCID. Since a keyword based search is not enough to avoid selecting an homonymous person, the chosen solution is the ORCID authentication method due to its reliability and integrity of allowing THOR applications to gain permission, through the unique user's credential, to read their ORCID records and accomplish a straightforward search of their ORCID ID value and other profile information, such as name, email and country. Additionally, discussions have begun on how data submission workflows can be coordinated with article submission workflows: a task of some complexity both technically and culturally.

PANGAEA, CERN and EMBL-EBI serve different research communities and therefore operate in different environments, and with different legacy workflows and technologies. It is important to note that the services delivered in this task have been integrated into live operational production systems. This therefore makes each of them a potential case study for other institutions in similar situations. This document describes the common experiences and challenges of the three institutions, as well as the specific and unique challenges and concerns each institution faced.

## 1. Introduction

ORCID provides a persistent digital identifier that distinguishes a researcher from every other researcher and, through integration in research workflows, supports automated linkages between researchers and their professional activities, ensuring that their work is recognised.

THOR is enabling the integration of ORCID in key data repositories researchers already use. Hence, when researchers submit datasets, data repositories automatically link submitted datasets to researchers' ORCID IDs.

We integrate ORCID into data submission workflows to include ORCID IDs in data records and to enable unambiguous credit mechanisms for the production of datasets, and propagate ORCID IDs into related services and infrastructures. Similar to the way in which articles can be uniquely associated with a person by virtue of inclusion in an ORCID works list, the integration work described below allows



datasets to be associated with a person in a manner that is unambiguous, easily propagated to ORCID and embedded in discipline-specific tools and workflows.

This document describes the experiences of three THOR partner institutions serving three distinct disciplines:

- The European Bioinformatics Institute (EMBL-EBI)
- The Publishing Network for Geoscientific & Environmental Data (PANGAEA)
- The European Organization for Nuclear Research (CERN)

These three partners developed services to integrate ORCID into their existing databases and workflows. Although the ultimate goal was the same, the specific situations of each institution naturally led to different requirements and work performed in support of this goal, leading in turn to a suite of complementary results: EMBL-EBI hosts several independently governed and separately funded databases, necessitating a centralised service for all databases to use; PANGAEA, which is not an ORCID member, must maintain its existing login system as well as offer the option to login with ORCID; CERN also maintains several databases but treats one of those databases as a central metadata store rather than developing a standalone centralised service.

The efforts are organised into two tasks, one for immediate development, and the other as the start of an ongoing effort that will run throughout the remainder of the THOR project.

The first task is to build common services that make use of harmonised metadata to integrate ORCID iDs into data records. Through outreach there must be engagement with database providers outside of THOR to encourage its adoption, avoid unnecessary effort duplication, and enable harmonisation and federation of workflows for researchers using different data repositories. Within this task:

- EMBL-EBI will develop tools to facilitate the incorporation of ORCID iDs into core life science databases, demonstrating their incorporation in at least two databases.
- CERN will implement ORCID authentication into the INSPIRE suggestion tool.
- PANGAEA will develop ORCID sign-in as an additional functionality to its existing login and enable the connection of PANGAEA user profiles with ORCID iDs.

The second task is an ongoing effort that demonstrates how ORCID iDs can be shared across workflows of article publication and data submission. The work will be synchronised with research and outreach activities that encourage and guide a wide range of data providers to adjust their data submission and ingest systems to push ORCID iDs to data management plans and datasets upon submission.

- EMBL-EBI will work with PLOS to enable sharing ORCID iDs for datasets in article-data publication workflows.
- CERN will focus on implementation of services for ORCID integration for several data submission systems and demonstrate that such data can be bi-directionally shared with the arXiv community platform and publishers.



- PANGAEA will integrate ORCID IDs into its data submission workflow, and collaborate with publishers in Earth and Environmental Science to further improve article-data publication and integration of ORCID IDs in such workflows. COPDESS (Coalition for Publishing Data in the Earth and Space Sciences) is a potential platform for such collaboration.

Initial work completed and plans for continuation are described in Section 6.

## 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 and research through the embedding of ORCID identifiers in key workflows, such as research profile maintenance, manuscript submission, grant application, and patent application.

ORCID provides two core functions: (1) a registry to obtain a unique identifier and manage a record of a researcher's own activities, and (2) APIs (Application Programming Interface) that support system-to-system communication and authentication. The ORCID Registry is available free of charge to individuals, who may obtain an ORCID identifier, manage their record of activities, and search for others in the Registry. Organisations may 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 is freely available via periodical 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 signed off on 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.

### 2.2 EMBL-EBI: The European Bioinformatics Institute

The European Bioinformatics Institute (EMBL-EBI)<sup>1</sup> is a centre for research and services in bioinformatics, and is part of the European Molecular Biology Laboratory (EMBL).

---

<sup>1</sup> <http://www.ebi.ac.uk/>

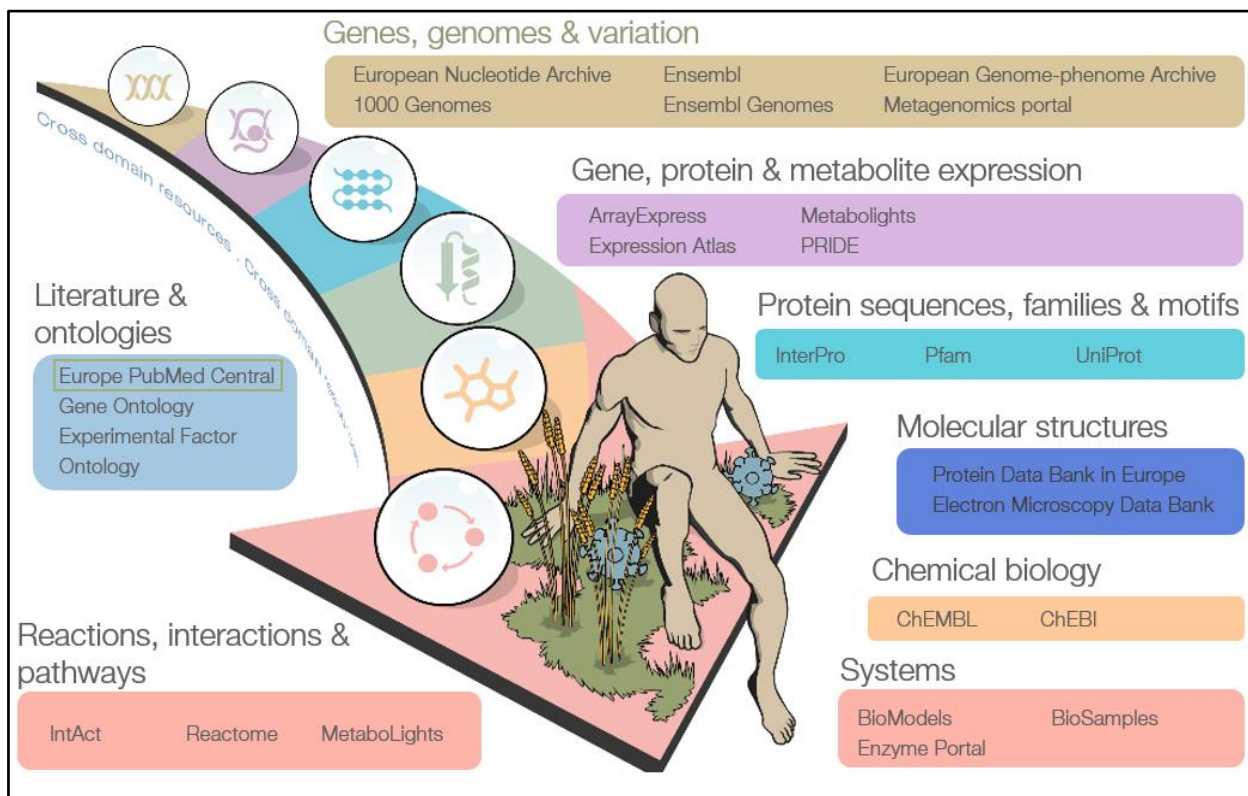


Figure 1: EBI core services

Figure 1 is a summary of EBI services. The core services are:

- **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 PubMed Central** – 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** – 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)



The Literature Services team integrated ORCID with Europe PMC, and manages EMBL's ORCID integration. Approximately 2.6 million articles now have links to ORCID IDs. With this experience as one of the core services at the EBI, the Literature Services team is engaging and collaborating with other services at the EBI to create reusable, harmonised workflows for databases to interact bidirectionally with the ORCID registry.

## 2.3 PANGAEA

PANGAEA, the Publishing Network for Geoscientific & Environmental Data<sup>2</sup>, is an Open Access library that archives, publishes and distributes geo-referenced data about climate variability, the marine environment and geological research.

We describe the workflows and services implemented in order to integrate ORCID in PANGAEA. The integration utilises the ORCID API to connect PANGAEA with the ORCID Registry using machine-to-machine communication. As PANGAEA is not an ORCID Member, it only has access to the ORCID Public API.

The ORCID–PANGAEA integration has the following aims:

1. Obtain the ORCID iD and other user profile data of ORCID-authenticated PANGAEA users
2. Provide PANGAEA users the option to sign in to PANGAEA using their ORCID account

To achieve these goals, PANGAEA extended its current sign-in functionality (with authentication based on its own user management system) with new functionality for authentication via ORCID. The ORCID “iD” button is included on the PANGAEA login form in order to trigger the ORCID authentication and authorisation workflow. PANGAEA user profiles are extended with an additional database field for the ORCID iD. Thus, PANGAEA caches the ORCID iD of users who connect their PANGAEA profile with their ORCID iD. For PANGAEA users who submit data, the cached ORCID iD is metadata automatically included in the data submission workflow.

## 2.4 CERN

CERN, 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**<sup>3,4</sup> – Our core service, a hub for literature in HEP. Built as an aggregator, it automatically harvests relevant literature from a set list of publishers. Literature relevancy is determined largely through automation with a human approval and suggestion layer.

---

<sup>2</sup> <http://www.pangaea.de/about/>

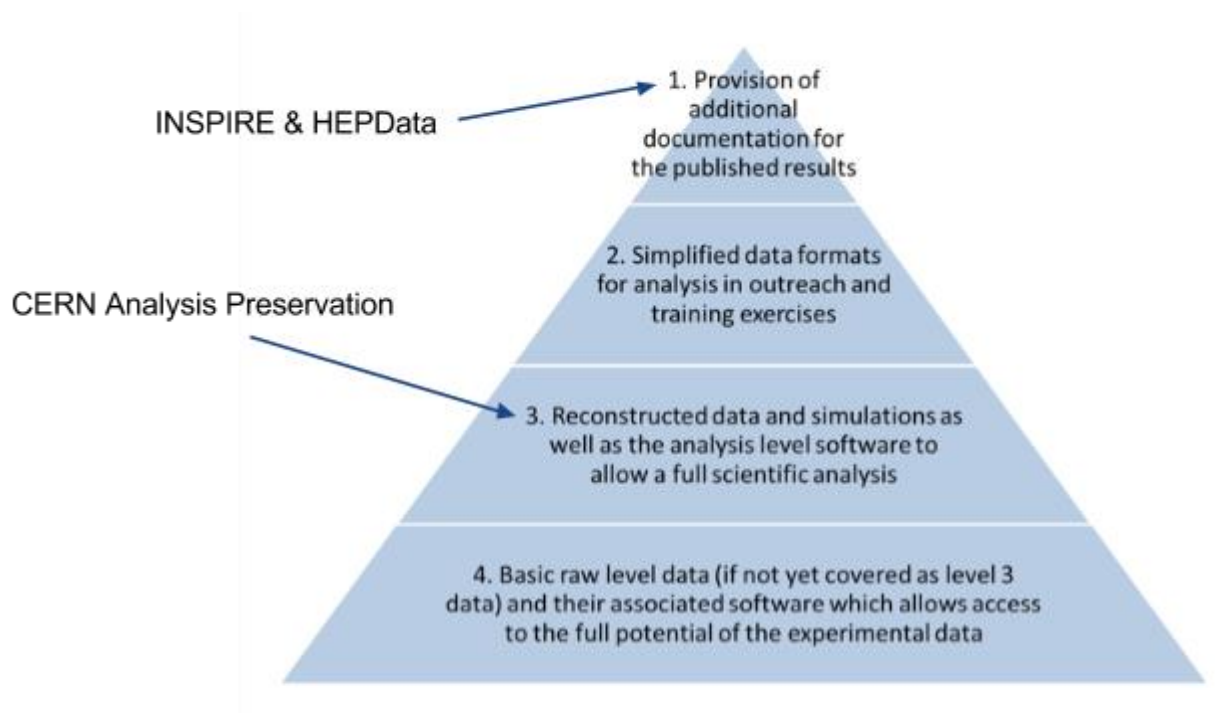
<sup>3</sup> <http://inspirehep.net/>

<sup>4</sup> <https://github.com/inspirehep>



- **HEPData**<sup>5,6</sup> – A collaboration between CERN and Durham University, explicitly for data that is supplemental to publications, e.g. final tables. Includes peer review workflows so that submitted data may be approved prior to release.
- **CERN Analysis Preservation**<sup>7,8</sup> – An internal tool for capturing the disparate files, metadata, and provenance associated with large-scale HEP analyses.

These services work in concert to address the data and research output needs at different stages of the progression of HEP research data from raw data to individual conclusions, as shown in Figure 2.



Pyramid of HEP data stages from: Herterich, P., & Dallmeier-Tiessen, S. (2016). Data Citation Services in the High-Energy Physics Community. *D-Lib Magazine*, 22(1/2). <http://doi.org/10.1045/january2016-herterich>

Figure 2: HEP data stages

<sup>5</sup> <https://hepdata.net/>

<sup>6</sup> <https://github.com/HEPData>

<sup>7</sup> <https://analysis-preservation-qa.cern.ch>

<sup>8</sup> <https://github.com/cernanalysispreservation>





### 3. Requirements for Integrating ORCID IDs into Data Records

EBI, PANGAEA and CERN all have a requirement to enable the users of their data services to authenticate via ORCID. In addition to providing a means of authenticating via a single authentication mechanism, this allows the institutions to retrieve useful information from ORCID for incorporation into their own services. It also allows them to push useful information back to ORCID.

While this may sound simple on the surface, the paths to this implementation taken by each institution have to fit with the institutions' development and service environments, as well as with the norms of the disciplines and communities they serve. These varied requirements and approaches are discussed in detail below, so that they may serve as examples for institutions with similar concerns.

#### 3.1 General Implementation of ORCID Authentication

In order to connect their services to ORCID, each institution made use of the ORCID RESTful API, which uses OAuth 2.0<sup>9</sup>. OAuth is an open standard that provides client applications with delegated access to a resource on behalf of the resource owner. With OAuth, a user can authorise a third-party system to access their account and resources without sharing their login information. Each institution has enabled the ORCID Public API via ORCID's Developer Tools<sup>10</sup> and registered an application for OAuth. ORCID then assigned a Client ID and Client Secret to each registered application. This is in accordance with the procedure outlined in the ORCID Public API documentation<sup>11,12</sup>, which can be consulted for further details.

Straight out of the box, the steps in the ORCID authentication workflow are as follows:

- A user clicks a link or button for authentication with ORCID, which triggers the **sign-in** (1) process.
- The institutional service redirects (2) the user to ORCID for user authentication and service **authorization**. If the user does not currently have an ORCID ID, she may register.
- Following successful authentication and authorisation, ORCID routes (3) the user back to the institutional service with an authorisation **code**.
- The institutional service exchanges (4) the code for an access **token**. This token is persistent and enables the institutional service to access the user's public ORCID metadata.
- ORCID then communicates (5) (some of) the user's public ORCID metadata to the institutional service. Of particular interest is the user's **ORCID ID**.

---

<sup>9</sup> <https://members.orcid.org/api/oauth2>

<sup>10</sup> <https://orcid.org/developer-tools>

<sup>11</sup> <http://members.orcid.org/api/introduction-orcid-public-api>

<sup>12</sup> <http://members.orcid.org/api/accessing-public-api>

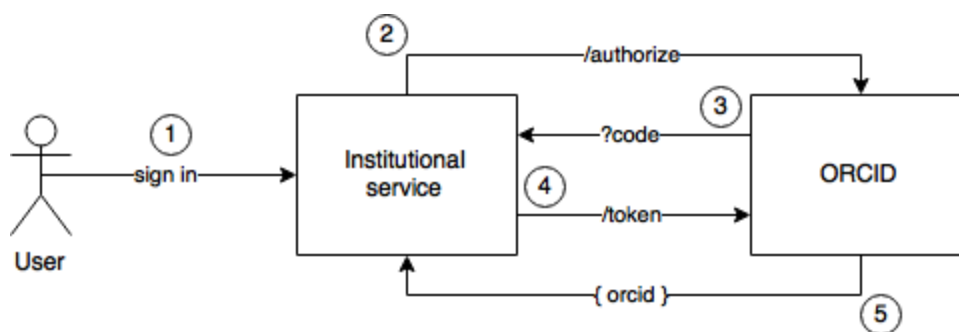


Figure 3: ORCID Authentication<sup>13</sup>.

Each institution described in this document implemented this basic workflow, with modifications or additions based on the specific needs of their services and/or users.

It should be noted that some potential functionality is discouraged by ORCID, as it is not within the spirit of user control over their academic record. For example, the ability to look up an individual by name and assign their ORCID to outputs is discouraged as this provides unauthenticated links and can lead to assignment errors. Bypassing the correct OAuth procedure by harvesting user credentials and posting them directly to the ORCID login service is also forbidden.

## 3.2 Institution-Specific Requirements

### 3.2.1 EMBL-EBI

As described above, the EBI hosts many life science databases that run with independent governance structures and with a variety of separate funding models; therefore, the EBI decided to provide a central service (API) that links to ORCID services for these databases to use. This will ensure that as ORCID is integrated into data resources at the EMBL-EBI there will be no duplication of effort (for example, every database processing the same piece of work), and ORCID integrations will have consistent UIs across different resources. Furthermore, the maintenance of a single service that interacts with ORCID can be managed effectively going forward. The central service needs to allow users, via the submission systems of different databases, to look up their ORCID iD after authentication on the ORCID website. It will return the ORCID iD with any publicly available ORCID metadata such as first name, family name, and affiliation.

### 3.2.2 PANGAEA

PANGAEA continues to maintain its own user management because the system builds on it. User management is needed to store information about users, including email and telephone numbers, and is used to track what a user does – and is allowed to do – on the website (for example, additional services,

<sup>13</sup> For more information on OAUTH2 protocol see: <https://members.orcid.org/api/oauth2>



data submission, download of restricted datasets). Consequently, to support ORCID sign-in PANGAEA must resolve the iD returned by ORCID to a PANGAEA user account. However, a user who is authenticated by ORCID may not be resolvable by PANGAEA. In order to be resolvable, PANGAEA users must have previously created a PANGAEA account *and* have connected her ORCID iD. In this case, PANGAEA can safely resolve the ORCID-authenticated user.

In contrast, an ORCID-authenticated user may be unresolvable by PANGAEA. This case occurs when the ORCID iD for an ORCID-authenticated user is unknown to PANGAEA. Thus, PANGAEA cannot safely resolve the user, and requests the ORCID-authenticated user first perform PANGAEA authentication and connect the ORCID iD. At this stage, the user may either sign up or sign in to PANGAEA, depending on whether or not the user has previously created a PANGAEA account.

An additional requirement for PANGAEA is to support the *connection* of an ORCID iD for a PANGAEA-authenticated user. In this case, an authenticated PANGAEA user decides to manually update her profile to include her ORCID iD. Disconnecting an ORCID account from a PANGAEA user profile is trivial as it amounts to simply removing any related ORCID information from the profile.

### 3.2.3 CERN

INSPIRE is CERN's main literature service, and the primary literature service for HEP. It is a largely automated aggregator with a human curation layer, and it is currently the main platform of the CERN scientific literature and research output system. It serves as the main metadata warehouse, maintaining both bibliographic and author information for (as far as possible) the entirety of HEP. Therefore, rather than develop a separate service and API to feed the various CERN literature and data services, the metadata from INSPIRE is used to inform the other products and services in CERN literature and data systems. It will be beneficial to INSPIRE – and thus, in turn, to the rest of the CERN ecosystem – to implement ORCID authentication as the sole means of authentication for INSPIRE services. In the past, users uploading datasets or making content or claim suggestions in INSPIRE could authenticate with arXiv or use the service as anonymous guests. Implementing ORCID as the only authentication mechanism will allow INSPIRE to require authentication, thus eliminating anonymous users and their added human oversight burden. Additionally, all authentication will be united under one commonly accepted author identification system, which is being adopted increasingly across publisher platforms, thus providing an additional benefit for users.

INSPIRE pushes and pulls works information to and from the ORCID profiles of those researchers for whom an ORCID iD is known. Moving to ORCID authentication will encourage others to connect their ORCID iDs with the INSPIRE profiles, and improve the quantity and quality of works information being pushed. HEPData, the platform for review and submission of supplemental publication data, pulls its metadata from INSPIRE. INSPIRE's implementation of ORCID authentication, and the subsequent updating of metadata with ORCID information, will in turn benefit HEPData. In addition, implementing ORCID authentication in HEPData itself would provide an alternative authentication method for the review workflow, although no metadata would be imported from or pushed to ORCID from HEPData directly.



## 4. Work Completed

### 4.1 EMBL-EBI

Middleware is a generic name that can be used to refer to software systems that mediate the communication between a client software application with another server software application. It consists of interfaces providing high level APIs that permit integration with diverse client applications by masking the heterogeneity of processes and simplifying the access to the server software application.

EBI has developed a middleware layer ("THOR ORCID Service") with an API library ("THOR ORCID Client") (see Figure 4). It can be incorporated into the client-side applications that enable EBI databases to seamlessly incorporate ORCID iDs into their resources.

The EBI's core services are provided by EBI databases (such as Metabolights, EMPIAR, PDBe, UniProt, among others) through their respective websites, which also accept the registration of individuals through their data submission forms. The integration of these forms with the THOR ORCID Client API enables the EBI databases to make available the link "Create / link an Open Researcher Contributor ID (ORCID)" in their data submission forms. With this link the users can authenticate at ORCID and authorise the obtaining of his ORCID iD value, first/family names and email from his profile.

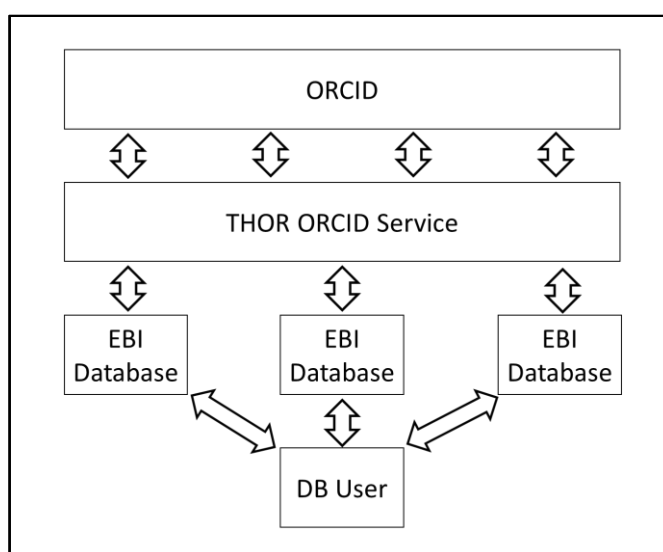


Figure 4: Overview of EBI databases accessing ORCID through middleware service

The steps required to achieve this are illustrated as below:

- Step 1: The user clicks the "Create/link an Open Researcher Contributor ID (ORCID)" link, and this will trigger the sign-in process. In turn, the JavaScript THOR ORCID Client connects to the THOR ORCID Service.



- Step 2 and 3: THOR ORCID Service redirects the user to the ORCID Registry for authentication and authorisation.
- Step 4 and 5: Once authenticated and authorised, the ORCID Registry will then direct the user back to the THOR ORCID Service with an authorisation code.
- Step 6 and 7: The THOR ORCID Service then exchanges the authorisation code for an access token.
- Step 8: The THOR ORCID Service will provide the database with the access token together with the corresponding ORCID ID.
- Step 9: The THOR ORCID Service will then return to the THOR ORCID Client with the user's information such as ORCID ID, first name, family name, country, and others from the public user profile, so that the page can be updated accordingly.

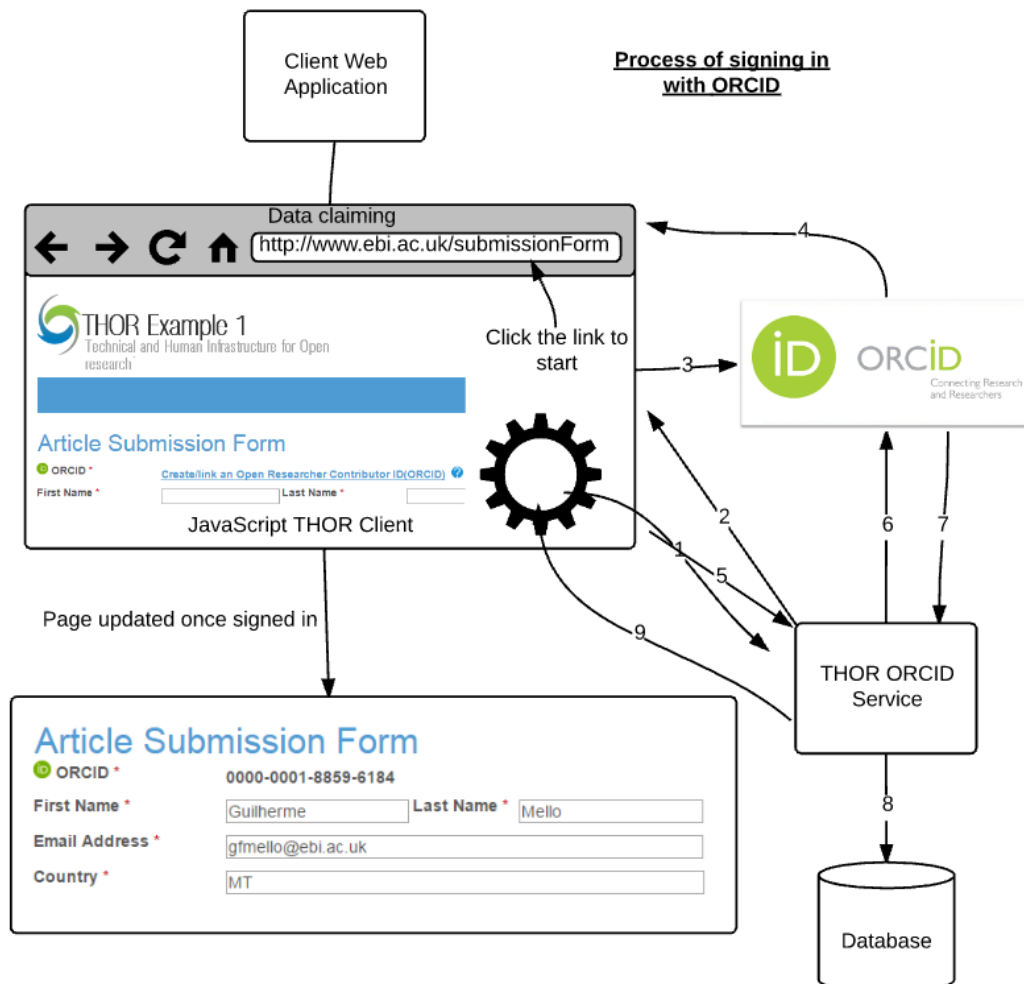


Figure 5: Sign in with EBI service



The THOR ORCID service is built as a Java application running on a Tomcat web server, hosted at the EMBL-EBI, using 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
- Client APIs over JavaScript and JQuery
- Transform the search results for data submission forms
- Manage the sign in process through OAuth 2.0 protocol
- Hibernate framework to retrieve access tokens from Oracle database
- Documentation

The Java API documentation for developers and additional implementation details links are provided in section 6.1.

## 4.2 PANGAEA

The interaction between PANGAEA and ORCID requires some programme logic, which PANGAEA implements in PHP and Java Web services for the frontend and backend, respectively.

PANGAEA has modified the login page for the relaunch of the website to include functionality to “Log in with ORCID iD”. This functionality is in addition to the existing login feature based on PANGAEA user management. The “Log in with ORCID iD” functionality is implemented by adding the ORCID “iD” button on the PANGAEA log-in page. Choosing to log in with ORCID triggers the workflow described above. Note that, at the time of writing, the new PANGAEA website was not yet released to the public. Public release is expected during mid-2016 and will come with a complete redesign and improved user experience.

PANGAEA modified the user profile page to include the ORCID iD, if connected. If a PANGAEA user has not connected her ORCID iD, the page displays the following text:

*“Your account is not yet connected to an ORCID iD. It is strongly recommended to connect your account, because ORCID provides a persistent digital identifier that distinguishes you from every other researcher and, through integration in key research workflows such as data set / manuscript / grant submission, supports automated linkages between you and your professional activities ensuring that your work is recognized. If you submit data to PANGAEA, we will link the data sets to your ORCID iD.”*

PANGAEA thus explicitly encourages its users to sign up at ORCID and/or connect their ORCID iD. If so desired, PANGAEA users can disconnect their ORCID iD by editing their user profile.

PANGAEA also modified its sign-up form to allow for the immediate connection of an ORCID iD. This is to enable the connection of ORCID iDs and PANGAEA user profiles as early as possible. The feature also comes with the added benefit that the “Full name” and “Institution/Affiliation” sign up form fields are automatically populated with information returned by ORCID. With a similar message as the one above



shown on the user profile page, the PANGAEA sign-up page also encourages users to connect their ORCID ID (and sign up for ORCID if they are not already registered).

PANGAEA developed a php-orcid library that implements some of the programme logic required for PANGAEA to implement the integration. The php-orcid library was published Open Source under Apache License Version 2 at the THOR project GitHub repository.<sup>14</sup> It is important to note that third party systems must redirect to ORCID for the OAuth authentication process. In other words, ORCID user credentials must be imputed by users into the ORCID sign in/register form, as opposed to the client application. Any attempt by any third party system to circumvent this redirection to ORCID is against ORCID guidelines and the OAuth Authorisation code flow specification. In practice, this means that a library cannot provide programme logic that returns an ORCID ID when given a username and password as input. Accordingly, the php-orcid library does not provide such programme logic. It only supports programme logic that returns an ORCID ID given the code returned by ORCID to a token request. (It is worth considering that future versions may also provide the ability to create the redirect URL.) The implementation of the ORCID authorise request remains the responsibility of PANGAEA or, more generally, the system that uses the php-orcid library. In addition to programme logic for the exchange of an ORCID code for a token, the php-orcid library also supports retrieving biographical and activity information about ORCID users.

It is important to note that ORCID IDs should under no circumstances be submitted as form fields (as text or hidden field) or HTTP GET/POST parameters, thus being transmitted from the client to the server. This practice enables the client to edit the submitted ORCID ID. PANGAEA stores the OAuth tokens and the authenticated ORCID ID in the browser session, secured by an extra one-time key transmitted as form-token<sup>15</sup> and using the “state” mechanism during the OAuth workflow<sup>16</sup>. It is therefore not possible to change the ORCID ID by manually overriding HTML form contents to take over another’s identity.

To support the caching of ORCID IDs, the PANGAEA database is extended with an additional attribute. The caching is necessary to automatically resolve ORCID-authenticated users to the corresponding PANGAEA profile. Caching of ORCID IDs is also useful to automatically assign the ORCID in user data submissions. With the introduced changes, authors of data submitted to PANGAEA have the possibility to link the submission with their ORCID ID.

PANGAEA has also developed an additional Java library called OrcidResolver, which supports the resolution of ORCID IDs of authors given their first name, last name, and a non-empty set of DOIs for “works” (possibly) claimed at ORCID (i.e. works added to an ORCID user profile either manually by the user or automatically by a trusted third-party system).

OrcidResolver is published Open Source under the Apache License Version 2. The library constructs a query with the given information, and requests the ORCID ID of the matching user from ORCID. The approach is reliable but does not guarantee that the returned ORCID ID is correct. With OrcidResolver,

---

<sup>14</sup> <https://github.com/thor-project/php-orcid/>

<sup>15</sup> [https://en.wikipedia.org/wiki/Cross-site\\_request\\_forgery#Synchronizer\\_token\\_pattern](https://en.wikipedia.org/wiki/Cross-site_request_forgery#Synchronizer_token_pattern)

<sup>16</sup> <https://tools.ietf.org/html/rfc6749#section-10.12>



relationships between users and ORCID iD are found algorithmically. It is important to underscore that the confidence in algorithmic approaches is weaker than approaches based on ORCID authentication. This programme code was originally developed by PANGAEA, and has been used to resolve ORCID iDs of the authors of its datasets. As the algorithm may be of interest to the wider community, PANGAEA has decided to refactor the original programme code into code suitable for a library that can be reused by others.

### 4.3 CERN

As part of a general site overhaul, CERN developed INSPIRE Labs: a testbed for new INSPIRE features. During the transition to the new version of INSPIRE, new features will be first hosted on Labs. At the time of writing, the INSPIRE frontend has not yet migrated to the new version, but many new features are being hosted on Labs and are linked from the legacy version of INSPIRE.

The transition to the new INSPIRE will also be a transition to a new and improved version of the underlying Invenio<sup>17</sup> platform. Invenio 3, the newest version of Invenio being developed by CERN IT in parallel with development of INSPIRE by CERN SIS, will include a module for OAuth support, allowing it to support authentication via third parties such as ORCID. As with other CERN services, the development of INSPIRE and Invenio is open, so code for the OAuth module is available on GitHub<sup>18</sup>.

As an interim solution, the INSPIRE team developed a Python wrapper library<sup>19</sup>, now maintained by ORCID, for the ORCID API to facilitate ORCID authentication in INSPIRE. This is currently implemented in INSPIRE, but it will be replaced by the Invenio OAuth module once the new version of INSPIRE is live. ORCID authentication has already been implemented for the content suggestion, correction and data submission services, which are now hosted on INSPIRE Labs. As stated previously, the move to ORCID authentication prohibits anonymous users, and helps to reduce the amount of human intervention necessary for claim and suggestion approval.

If a user has previously added her ORCID iD to her INSPIRE author profile and authenticated from her INSPIRE author profile page, INSPIRE pushes works information to her ORCID profile. At this time, it is only appending works to the profile, not correcting existing metadata.

This functionality is currently separate from ORCID authentication on INSPIRELabs, which focuses on submissions and suggestions for new records. There, ORCID authentication does not yet provide ORCID iDs to the INSPIRE author record automatically. If a user logs in via ORCID to make a submission or suggestion, INSPIRE Labs collects their name and ORCID iD as a means of identification, but at present this information is not associated with an existing INSPIRE author profile automatically. This is due to change with the upgrade of the claiming service in summer 2016, and thus will be detailed as part of forthcoming THOR work on data claiming services.

---

<sup>17</sup> <http://invenio-software.org/>

<sup>18</sup> <https://github.com/inveniosoftware/invenio-oauthclient>

<sup>19</sup> <https://github.com/ORCID/python-orcid>



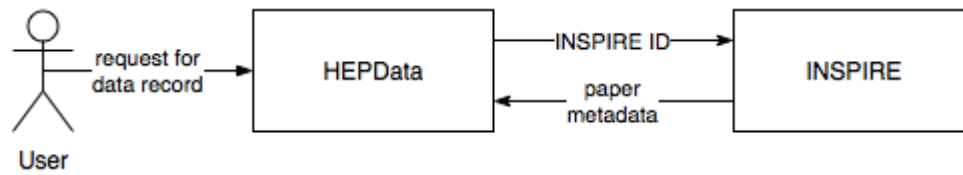


Figure 6: HEPData retrieves metadata from INSPIRE based on the paper’s INSPIRE ID associated with the record

INSPIRE also pulls information from ORCID. Non-INSPIRE works that are discovered in ORCID profiles are displayed on the External tab of the INSPIRE author profile.

HEPData pulls all of its bibliographic and author metadata from INSPIRE (a data set must have an associated INSPIRE ID for a publication in order to be released), and does not store any author metadata of its own (see Figure 6). Any author metadata enhancements must happen on INSPIRE. If the INSPIRE record includes an ORCID iD, that will in turn be associated with the HEPData record by virtue of HEPData re-using INSPIRE metadata. In light of this, ORCID authentication is in place in HEPData as an alternative login for HEPData’s review functions, but it is not currently used for metadata enhancement purposes.

CERN Analysis Preservation – the newest CERN data service – has made significant progress over the course of its first year, both in terms of general development of the frontend and backend, as well as building connections to fetch research content from databases internal to the individual CERN experimental collaborations. It is now available as a first prototype.

ORCID authentication has been implemented in CERN Analysis Preservation, but it has not yet been switched on for users; this will remain so until a solution to the human challenge can be determined. CERN Analysis Preservation is an internal platform, not open to those outside the experimental collaborations, and so the service must follow the experiments’ detailed access regimes. CERN Analysis Preservation records will be populated with ORCID iDs once the various CERN experimental collaborator databases from which CERN Analysis Preservation harvests are populated with ORCID iDs. Therefore, the service currently focuses on CERN authentication, while the CERN SIS team works to foster the uptake and wider usage of ORCID iDs within the immediate CERN community – a task that will benefit from cooperation with the outreach component of THOR.

The ongoing development of CERN Analysis Preservation raises other interesting challenges for THOR (for example, dynamic data citation) that will be addressed in detail in later project outputs.



## 5. Results

This section presents a series of screenshots that demonstrate how EMBL-EBI, PANGAEA, and CERN have integrated ORCID into their services and databases.

### 5.1 EMBL-EBI

EBI has developed a service with an API that any number of databases at EBI can use to interface with the ORCID registry to verify user identity and populate existing data submission forms with data from their ORCID profile. Special emphasis was put on the simplicity of integrating existing web forms with the API. The result is that it takes just a few lines of code to embed the API. Two databases, Metaboblights and EMPIAR, have already integrated this into their workflows.

In addition, EBI are actively engaging with other life science databases (BioStudies, PDBe, the European Nucleotide Archive, IntAct, PRIDE proteomics database and ArrayExpress), and expect these to start using the API over the next months and years, depending on agreements with their collaborators, governance structures, and production development plans.

1. The databases that have been integrated with the API are going to automatically display a link that enables the user to authenticate at ORCID (see Figure 7).

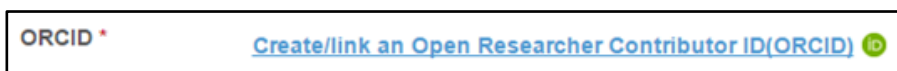


Figure 7: Link generated by the API for the user authentication

2. ORCID authentication screen for EMBL requesting permission to read and write to a user's ORCID record (see Figure 8).

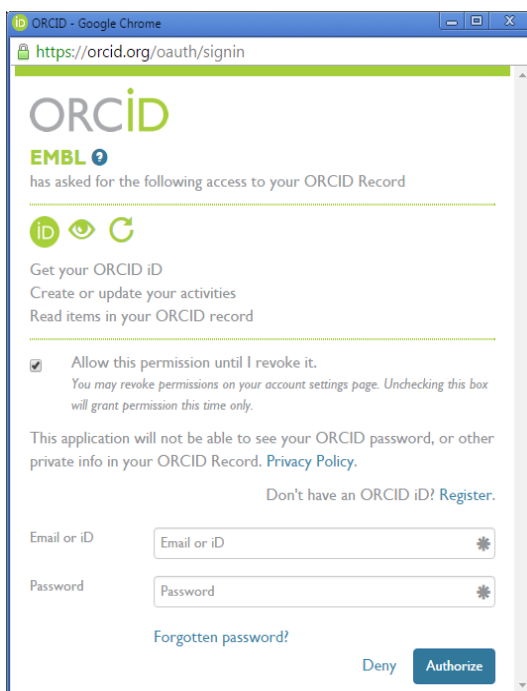


Figure 8: ORCID authentication popup



3. Use Cases of the EMBL-EBI ORCID-integration APIs in data resources.

- (a) **Metabolights Database.** This database receives submissions of Metabolomics experiments and derived information. The database is cross-species, cross-technique and covers metabolite structures and their reference spectra as well as their biological roles, locations and concentrations, and experimental data from metabolic experiments.

Figure 9 and Figure 10 show Metabolights registration forms with integrated ORCID authentication. In the first figure a link is displayed for the user to authenticate within ORCID, whereas in the second the form fields are filled automatically after the user authenticates at ORCID.

The screenshot shows the MetaboLights registration page. The header includes the EMBL-EBI logo and navigation links. The main heading is "MetaboLights" with a search bar. Below the navigation bar, the page title is "MetaboLights > Create a new account". The main content area contains the text "Enter your details to request a new account for MetaboLights." followed by a form with the following fields: "Email address\*", "ORCID:" (with a link "Create/link an Open Researcher Contributor ID(ORCID)"), "First name(s)\*", "Last name\*", "Password\*", "Password (repeated)\*", "Country\*" (dropdown), "Affiliation/Institution\*", and "Web address for affiliation/institution (URL)\*". There is also a checkbox for "I understand that all public data is freely available by any individual and for any purpose." and "Create" and "Cancel" buttons.

Figure 9: Metabolights registration form integrated with ORCID authentication

The screenshot shows the same MetaboLights registration page, but now the ORCID field is pre-filled with "0000-0002-9829-091X", the first name field with "Guilherme", and the last name field with "Mello". All other fields and the checkbox remain the same as in Figure 9.

Figure 10: Metabolights registration form after user authenticates on ORCID



- (b) **EMPIAR Database.** Database of electron microscopy images in structural biology submission page. Figure 11 and Figure 12 show EMPIAR registration forms with integrated ORCID authentication. In the first figure a link is displayed for the user to authenticate within ORCID, whereas in the second the form fields are filled automatically after the user authenticates at ORCID.

The screenshot shows the EMPIAR registration page. The header includes the EMBL-EBI logo, the EMPIAR logo, and the text "Electron Microscopy Pilot Image Archive". Navigation links for "Services", "Research", "Training", and "About us" are present. A search bar is in the top right. Below the header, there are links for "EMPIAR home", "Deposition", "Annotation", "FAQ", and "About EMPIAR", along with "Share" and "Feedback" buttons. On the left, a sidebar contains "EMPIAR deposition system" with links for "Deposition manual" and "Log in". The main content area is titled "Register with the EMPIAR deposition system" and includes a link "Create/link an Open Researcher Contributor ID(ORCID)". The registration form has the following fields:

ORCID	<input type="text"/>
First name *	<input type="text"/>
Middle name	<input type="text"/>
Last name *	<input type="text"/>
Username *	<input type="text"/>
Password *	<input type="password"/>
Repeat your password. *	<input type="password"/>
Email *	<input type="text"/>
Repeat your email *	<input type="text"/>

Figure 11: EMPIAR registration form integrated with ORCID

This screenshot shows the same EMPIAR registration page as Figure 11, but with the form fields populated with user information:

ORCID	0000-0002-9829-091X Example: 0000-0001-7663-9028
First name *	Guilherme Example: John
Middle name	<input type="text"/> Example: William
Last name *	Mello Example: Smith
Username *	<input type="text"/>
Password *	<input type="password"/>
Repeat your password. *	<input type="password"/>
Email *	<input type="text"/> Example: help@example.com
Repeat your email *	<input type="text"/> Example: help@example.com

Figure 12: EMPIAR registration form after user authenticates on ORCID



## 5.2 PANGAEA

PANGAEA users are now able to connect their ORCID iD with their PANGAEA user profile. With this, PANGAEA further improves its support for associating users with ORCID iDs. PANGAEA now has a trusted mechanism to identify authors uniquely, and to attribute PANGAEA data submissions to authors unambiguously. Identifying authors by their ORCID iD is a crucial aspect. The actual first name, last name, email and other information about an author are attributes of the particular ORCID iD. These attributes may change without interfering with data publication attribution. For instance, a researcher's family name may change over time but her ORCID iD stays the same, enabling accurate attribution of her work at any time. Technically, PANGAEA could delegate attribute value management entirely to ORCID. The trusted connection of ORCID iD makes such behaviour in principle feasible, provided the ORCID user sets her privacy settings accordingly. Whether or not it is desirable is another question. For instance, the approach poses some technical challenges, as attribute values may have to be cached and synchronised between systems in order to retain acceptable performance.

A second result is the possibility for PANGAEA users to choose to sign in to PANGAEA via ORCID. Especially for users who regularly use ORCID, this feature can be attractive as they may be signed into PANGAEA automatically as a result. The ORCID-PANGAEA integration relaxes the requirement of having to remember user credentials for both the PANGAEA and the ORCID accounts. Finally, the sign-in integration follows a trend of supporting third-party sign in at portals. Many examples can be found, including some that support ORCID sign in, such as publons.com. Supporting the connection of ORCID iD in author profiles is also a notable trend in manuscript and peer-review tracking systems, such as Editorial Manager. The possibility of capturing ORCID iDs of authors at manuscript submission is shared with PANGAEA, as is the possibility of capturing ORCID iDs of authors at data submission.

Figure 13 is an example screenshot<sup>20</sup> of the PANGAEA login page with “iD” button. Users can thus choose to log in using PANGAEA credentials or ORCID.

---

<sup>20</sup> The ORCID-PANGAEA integration is implemented in the next version of the PANGAEA website (<https://www.pangaea.de>); the screenshots shown here are for the new version. At the time of writing, the new version was not yet released to the public. Public release is expected during Q2 2016.



**PANGAEA.**  
Data Publisher for Earth & Environmental Science

SEARCH SUBMIT ABOUT CONTACT

### PANGAEA Log in

**Login is only required for access to data under moratorium or for submitting new data sets.**

*Most of the data are freely available and can be used under the terms of the license mentioned on the data set description. A few password protected data sets are under moratorium from ongoing projects. The description of each data set is always visible and includes the principle investigator (PI) who may be asked for access.*

You can [sign up for a user account at PANGAEA here](#). This account can be used to access more advanced services (like our data warehouse), access your own data under moratorium, or submit data using the issue tracker.

**Log in with ORCID iD**

Your account must already be connected to an **ORCID iD** for this to work. **ORCID** provides a persistent digital identifier that distinguishes you from other researchers. If your account is not yet connected, please use the [username/password log in](#) below and connect the accounts from the profile page after logging in.

Log in with ORCID iD  Keep logged in on this computer

**Log in with username and password**

User Name / E-mail:

Password:

Lost password?   Keep logged in on this computer

PANGAEA IS HOSTED BY  
Alfred Wegener Institute, Helmholtz Center for Polar and Marine Research (AWI)  
Center for Marine Environmental Sciences (MARUM)

PANGAEA IS MEMBER OF

Follow us

THE SYSTEM IS SUPPORTED BY  
The European Commission, Research  
Federal Ministry of Education and Research (BMBF)

Figure 13: PANGAEA login screen with the new option to sign in with ORCID iD in addition to the existing login with PANGAEA user credentials.



Choosing to log in using ORCID triggers the ORCID integration workflow described earlier. Specifically, the user is redirected to ORCID which will present him with the form shown in Figure 14.

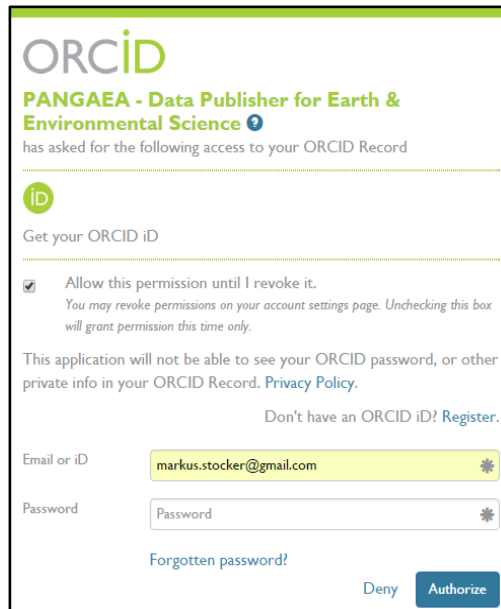


Figure 14: ORCID authentication form for user authentication and request to authorise PANGAEA to obtain the user’s ORCID iD

Following ORCID authentication and authorisation, PANGAEA displays the user as logged in. Figure 15 shows PANGAEA’s home page and highlights “Markus Stocker” as the logged-in user (upper right corner). Naturally, this behaviour is independent from the log-in path chosen by the user, i.e. log in via ORCID or PANGAEA.

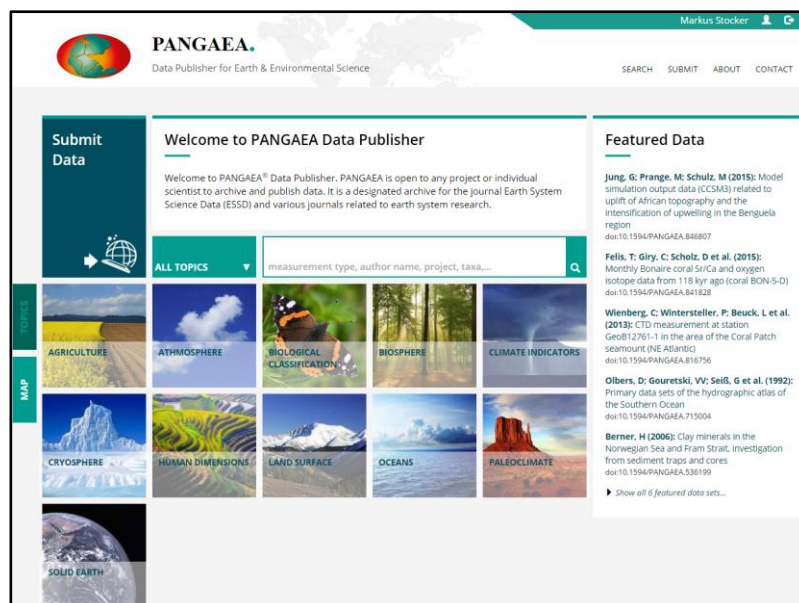


Figure 15: PANGAEA homepage showing user “Markus Stocker” logged into PANGAEA following ORCID user authentication and service authorisation (upper-right corner)



**PANGAEA.**  
Data Publisher for Earth & Environmental Science

Markus Stocker

SEARCH SUBMIT ABOUT CONTACT

### User Profile: mstocker

Hallo Markus Stocker,

On this page you have access to your PANGAEA user profile. You are logged in as **mstocker** with the following options:

- [Change password](#)
- [Edit user profile](#) (full name, e-mail, institution, phone)
- [Review our privacy policy](#)
- [Log out](#) and return to PANGAEA home page

Your account is connected to the following **ORCID iD**:  
<http://orcid.org/0000-0001-5492-3212>

Figure 16: PANGAEA user profile information with possibility of changing the password and edit the user profile. Now users are also informed about the connected ORCID iD. The page informs, and strongly encourages users without a connected ORCID iD about the possibility of creating an ORCID account and connecting the iD in PANGAEA.

### Sign Up

You can sign up for a user account using this form. This account can be used to access more advanced services (like our data warehouse) or access data under moratorium, or submit data using the issue tracker.

*Most of the data are freely available and can be used under the terms of the license mentioned on the data set description. A few password protected data sets are under moratorium from ongoing projects. The description of each data set is always visible and includes the principle investigator (PI) who may be asked for access.*

Connected ORCID iD: <http://orcid.org/0000-0002-1900-4162>

User name\*:

E-mail address\*:

Reenter e-mail address\*:


Password\*:

Reenter password\*:

Full name\*:

Institution/Affiliation:

Phone:

Captcha\*:  I'm not a robot  reCAPTCHA  
Privacy - Terms

Yes, I have read the [privacy policy](#) of PANGAEA (\* denotes a required field in this form).

Figure 17: PANGAEA registration form whereby the new user “Uwe Schindler” has connected his ORCID iD. Some of the required form fields are automatically populated with information obtained from ORCID about “Uwe Schindler”.





As shown in Figure 16, in addition to allowing the user to change password and edit the user profile, the PANGAEA user profile overview page now also informs users about the connected ORCID ID.

Figure 17 shows the “Sign Up” form used to create new PANGAEA accounts. The user “Uwe Schindler” has already connected his ORCID ID. As shown, Uwe Schindler’s full name and affiliation are populated *automatically* with information obtained from ORCID. PANGAEA still requires the user to provide a username and password because accounts can be disconnected from ORCID, so “backup credentials” need to be supplied. Unfortunately, in most cases a verified e-mail address cannot be obtained from ORCID because the vast majority of ORCID users choose not to share emails (the default setting). Hence, after account signup the user must run through the usual e-mail verification steps.

### 5.3 CERN

ORCID authentication is currently implemented on INSPIRE Labs, the testing ground for new INSPIRE features, in support of the content suggestion and data submission features demonstrated in Figure 18 and Figure 19. The page with the ORCID login button is reached by clicking Sign In (upper right) on any page of INSPIRE Labs. All links on legacy INSPIRE that once pointed to addition, suggestion, or submission forms now route to INSPIRE Labs.

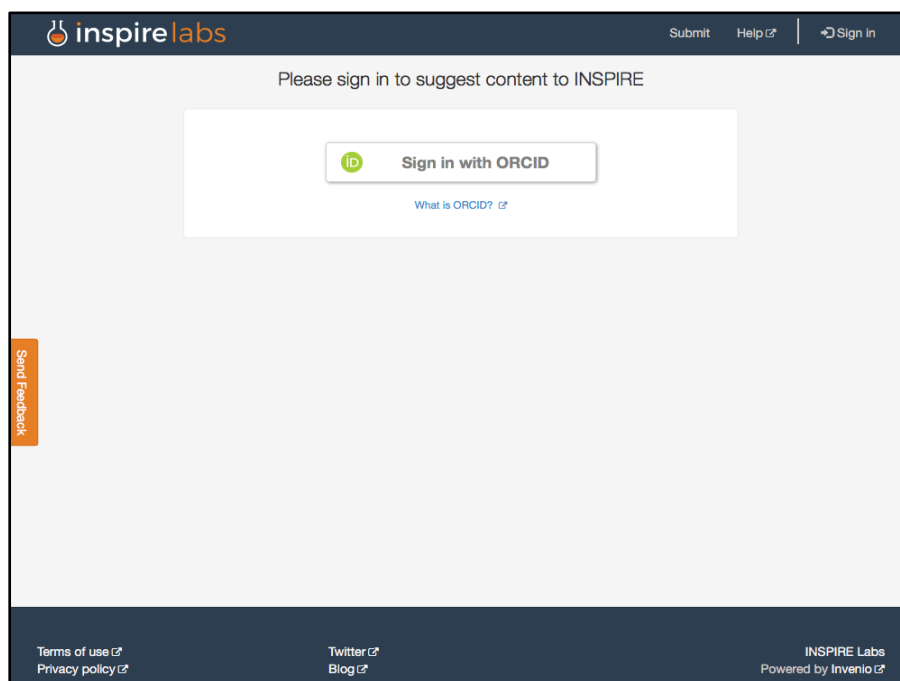


Figure 18: INSPIRE Labs sign-in screen. ORCID sign-in is necessary to suggest content or to submit data.

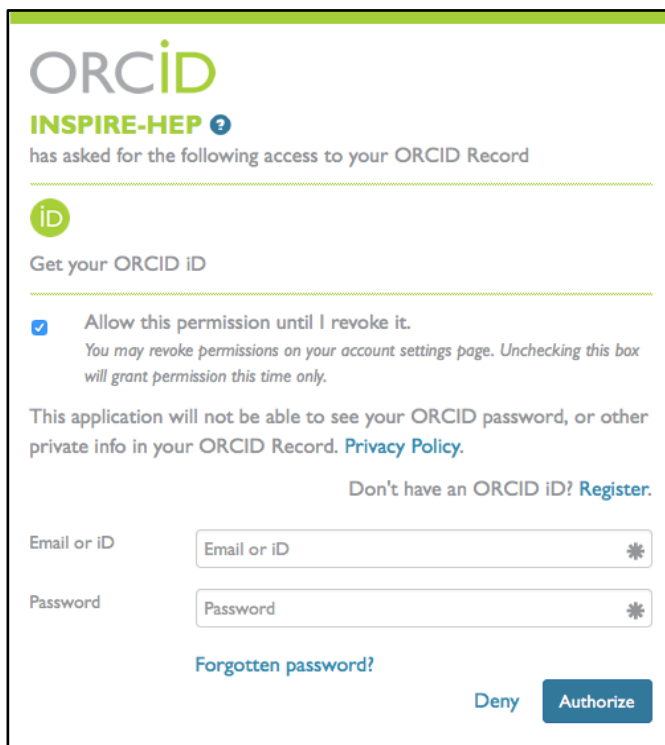


Figure 19: ORCID authentication screen

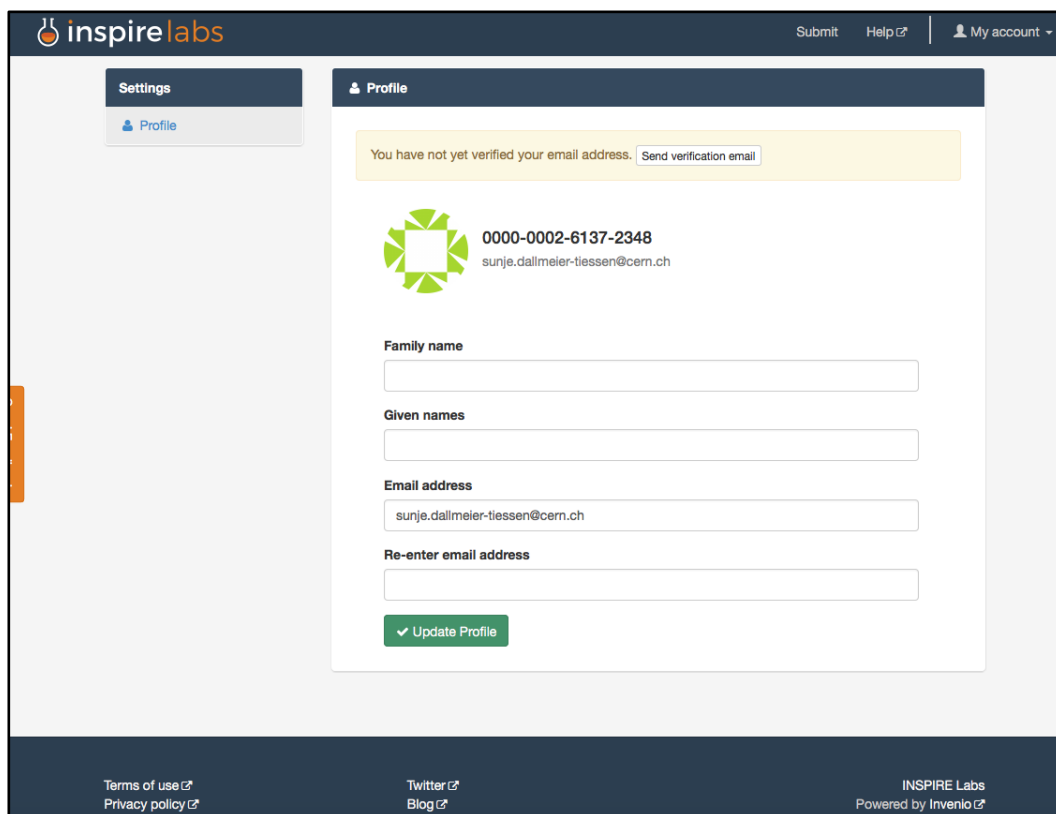


Figure 20: INSPIRE Labs profile page, showing the connected ORCID iD after authentication



As Figure 20 shows, similar to the PANGAEA workflow, users are asked to supply their email address when using ORCID authentication on INSPIRE for the first time. This enables INSPIRE to easily get in touch with the submitters in case there are questions about the dataset or suggestion submitted.

Figure 21 shows that ORCID IDs are included as part of author identity information within a user's public INSPIRE profile.

**inSPIRE** HEP

Welcome to INSPIRE, the High Energy Physics information system. Please direct questions, comments or concerns to [feedback@inspirehep.net](mailto:feedback@inspirehep.net).

HEP :: HEP-NAMES :: INSTITUTIONS :: CONFERENCES :: JOBS :: EXPERIMENTS :: JOURNALS :: HELP

### Cranmer, Kyle S.

Profile Name  Search

2016-04-25 16:13:09

[View Profile](#) [Manage Profile](#) [Manage Publications](#) [Help](#)

**PERSONAL INFORMATION**

**PERSONAL DETAILS (HepNames)**

Name: Kyle S. Cranmer  
 Current Institution: New York U.  
 E-mail: [cranmer@cern.ch](mailto:cranmer@cern.ch)  
 Links: <http://physics.as.nyu.edu/obje...>  
<http://twitter.com/KyleCranmer...>  
<http://hepandpractice.org/>  
 Fields: 0030  
 Experiment: LHC-LHC-ATLAS, LERN-LEP-ALEPH  
 Identifiers: BAI: K.S.Cranmer.1, INSPIRE: INSPIRE-00074922, ORCID: 0000-0002-5769-7094, ARXIV: cranmer\_k\_1  
 Institution: Rice U.  
 Period: 2007, 2005-2007, 1999-2005 (PHD), 1995-1999 (UG)

**PUBLICATIONS AND OUTPUT**

Publications Datasets External

1. Measurements of the charge asymmetry in top-quark pair production in the dilepton final state at  $\sqrt{s} = 8$  TeV with the ATLAS detector
2. Measurements of  $Z\gamma$  and  $Z\gamma\gamma$  production in  $pp$  collisions at  $\sqrt{s} = 8$  TeV with the ATLAS detector
3. Search for metastable heavy charged particles with large ionization energy loss in  $pp$  collisions at  $\sqrt{s} = 13$  TeV using the ATLAS experiment
4. Study of the rare decays of  $B_s^0$  and  $B^0$  into muon pairs from data collected during the LHC Run 1 with the ATLAS detector
5. Search for the Standard Model Higgs boson decaying into  $b\bar{b}$  produced in association with top quarks decaying hadronically in  $pp$  collisions at  $\sqrt{s} = 8$  TeV with the ATLAS detector
6. Measurement of fiducial differential cross sections of gluon-fusion production of Higgs bosons decaying to  $W^+W^- \rightarrow \mu\nu\mu\bar{\nu}$  with the ATLAS detector at  $\sqrt{s} = 8$  TeV
7. Search for new phenomena in events with a photon and missing transverse momentum in  $pp$  collisions at  $\sqrt{s} = 13$  TeV with the ATLAS detector
8. Measurement of  $W^+$  and  $Z$ -boson production cross sections in  $pp$  collisions at  $\sqrt{s} = 13$  TeV with the ATLAS detector
9. Search for charged Higgs bosons produced in association with a top quark and decaying via  $H^{\pm} \rightarrow \tau\nu$  using  $pp$  collision data recorded at  $\sqrt{s} = 13$  TeV by the ATLAS detector

**Co-Authors**

B.Mellado.1 (13)  
 W.Quayle.1 (11)  
 C.T.Potter.1 (9)  
 I.Aracena.1 (9)  
 M.Walters.1 (9)  
 S.L.Wu.1 (9)  
 A.T.Watson.1 (8)  
 B.Vachon.1 (8)  
 C.Santamarina-Ross.1 (8)  
 S.H.Robertson.1 (8)  
 more

**Papers**

	All papers	Single authored
All papers	696	11
Book	0	0
ConferencePaper	34	9
Introductory	0	0
Lectures	0	0
Published	591	4
Review	5	0
Thesis	1	1
Proceedings	0	0

**STATS**

**Citations Summary**

696 papers found, 688 of them citeable (published or arXiv)

	Citeable papers	Published only
Number of papers analyzed:	688	591
Number of citations:	63205	60097
Citations per paper (average):	91.9	101.7
$h_{\text{HEP}}$ index [?]	112	110

Breakdown of papers by citations:

	Citeable papers	Published only
Renowned papers (500+)	16	15
Famous papers (250-499)	17	16
Very well-known papers (100-249)	105	103
Well-known papers (50-99)	159	156
Known papers (10-49)	259	250
Less known papers (1-9)	100	50
Unknown papers (0)	32	1

[Click here to view statistics without self-citations or RPP](#)

Warning: The citations count should be interpreted with great care. [Read the fine print](#)

Publication Graph

Figure 21: ORCID ID included in author's public INSPIRE profile



## 6. Challenges and Lessons Learned

### 6.1 Open Development and Code Sharing

The opportunity to develop a common code base for interacting with the ORCID infrastructure was discussed, but proved complex due to the different contexts and infrastructures existing at the three participating institutions. For example, PANGAEA receives access to the ORCID API partially through DataCite and has a single major database service; EMBL-EBI has direct access to the APIs, but needed to integrate ORCID IDs into many EBI-hosted databases. Our remit is to deliver real services in production environments; therefore, pragmatic decisions on technology, expertise, and timing of developments were required.

Table 1 lists the technologies used in the three different production environments, links to source code, and documentation (also available via the THOR Knowledge Hub<sup>21</sup>). The open availability of source code and documentation will enable others to reuse it in a variety of contexts. The discussion of long-term sustainability and availability of source code, including where that code will reside after THOR, will be addressed as part of the project's overall sustainability work.

Table 1: Source code and documentation information for developers

Institute	Technologies Used	Link to Code	Link to documentation for ORCID Integration
EMBL-EBI	Java, JavaScript	Source code: <a href="https://github.com/thor-project/ebi">https://github.com/thor-project/ebi</a>	Documents for developers: <a href="http://www.ebi.ac.uk/europepmc/thor/resources/doc/">http://www.ebi.ac.uk/europepmc/thor/resources/doc/</a>  Client integration details: <a href="http://www.ebi.ac.uk/europepmc/thor/resources/DocumentForOrcidAuthorization.pdf">http://www.ebi.ac.uk/europepmc/thor/resources/DocumentForOrcidAuthorization.pdf</a>
CERN	Python, JavaScript	Python wrapper library for ORCID API: <a href="https://github.com/ORCID/python-orcid">https://github.com/ORCID/python-orcid</a>  Invenio OAuth client module: <a href="https://github.com/inveniosoftware/invenio-oauthclient">https://github.com/inveniosoftware/invenio-oauthclient</a>	Invenio OAuth client module: <a href="http://pythonhosted.org/invenio-oauthclient">http://pythonhosted.org/invenio-oauthclient</a>
PANGAEA	Java, PHP	Component to resolve ORCID ID based on author name and DOI list: <a href="https://github.com/thor-project/OrcidResolver">https://github.com/thor-project/OrcidResolver</a>  PHP implementation of login and account linking: <a href="https://github.com/thor-project/php-orcid">https://github.com/thor-project/php-orcid</a>	Documentation in source code (Javadoc's, PHP docs)

<sup>21</sup> <https://project-thor.readme.io/>



The results listed in Section 5 show that technical deliverables have been achieved. However, it is a longer-term, non-technical goal to stimulate uptake and acceptance of ORCID IDs for data, and the recognition of data as a first-class research object. The social aspects of this task are non-trivial, and so will take significantly longer to gain traction than the technical deliverables. The work of THOR outreach will be critical to build on this work.

## 6.2 Considerations on Authorship and Submission Workflows

The current implementation does not address the capture of ORCID IDs for co-authors in data submissions. The preferred method to include ORCID IDs in data (or article) records is via authentication, which by definition can only be done by one person – the submitting author.

The outcomes of a forthcoming THOR report on dataset claiming services and workflows will address this matter by allowing data submitters to claim data records to their ORCID profiles retrospectively, after the data have been released. It may also be possible to consider the approach employed by journals, namely to request email addresses of co-authors on submission, email the co-authors to inform them that data has been submitted on their behalf, and suggest that they validate the submission by authenticating the record in the submission system through the addition of their ORCID iD. It should be noted that for many databases, this would require significant changes to their author/submitter-handling pipelines. Furthermore, to be effective, it would require a social shift around recognising data as a first-class research object: databases and submitters would need to afford the same status to the authorship of datasets as they do article authorship, and make both social and technical adjustments as required.

Sharing of author and submitter data between articles and data workflows, such as ORCID IDs, is another possibility, although there are concerns here regarding privacy and accuracy, at least in the life sciences. To do this algorithmically, certain assumptions have to be made (for example, the same names on author and submitter lists being the same person). Life sciences authorship lists can number hundreds of people; it is not at all certain that those same people should be those acknowledged as data creators. Life science experimental outputs can be complex and enormous (as, for example, in the case of the 1,000 genomes project). These larger projects have greater potential for errors from, for example, algorithmic allocation of ORCID IDs in either direction.

PANGAEA has been successfully using this approach since 2013 and has now released it as an Open Source component, called the “OrcidResolver”. The approach takes the author name and all related dataset and article DOIs known to PANGAEA, and searches ORCID for a unique hit. Most datasets at PANGAEA are linked directly to articles (for example, as supplementary data to articles). Furthermore, these links are manually curated at PANGAEA. Curating the link between datasets and articles has proven sufficient for accurate ORCID ID resolution. In fact, PANGAEA has validated the 5% resolved ORCID IDs for its user base (40,000) and has, throughout the years, not found an erroneously resolved ORCID ID.



One of the subtasks underlying the efforts described in this document was to explore workflows around article and data submission. Some initial discussions and documents have been generated, but this is a complex task with many stakeholders involved. In order to address it in a comprehensive manner, it needs to operate on live systems that are handling hundreds of data and article submissions per day. However, these discussions have identified five scenarios that represent the dynamics between article and data submission, the distinctions between which make it easier to serve the individual use cases.

These are:

1. Data is deposited and the repository issues an accession number/DOI but the record is held until the article is published. An article can cite the data's DOI or accession number, but there are gaps, for instance, in the repository receiving notification of publication from the journal.
2. Data is submitted on article submission.
3. Data is submitted only once the paper is accepted.
4. Data is deposited after the article has been published. It is difficult in this case to link to the data from the article as the scientific record (article) is static.
5. Data is publicly available prior to the publication (i.e. available data has been reused).

PLOS has drafted information on the specifics of its data submission workflows, and a hands-on workshop will be held to refine thinking about what could be done in each of the above scenarios. CERN has been revising its article-data link exposure, and has been liaising with publishers such as Elsevier to implement workflows that enable seamless bi-directional article-data linking during the publishing workflow.

This task in particular is expected to also involve the research arm of THOR and potentially be the focus of a wider outreach workshop.

## 7. Conclusion

While the three institutions involved in this work differed in context and execution, the development work to enable ORCID authentication in each of our systems was implemented with relative ease. The real challenges come with the human factors involved in research uptake, both of ORCID IDs and of new systems and services. This supports THOR's mission of addressing both technical and human infrastructure in concert, and further emphasises the need for THOR's outreach work to communicate the value of the project's development efforts.

As this work was in progress, DataCite released DataCite Event Data<sup>22</sup>, a tool for tracking citation and update events for DataCite DOIs. This is an important step towards better article-data linking. DataCite Event Data will be beneficial for the ongoing journal integration work begun as part of the efforts described in this document, and it will be integrated into the THOR development roadmap.

---

<sup>22</sup> <https://eventdata.datacite.org/>



## Appendix A: 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. <a href="http://home.cern/">http://home.cern/</a>
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 <a href="https://www.datacite.org">https://www.datacite.org</a>
DOI	Digital Object Identifier
EC	European Commission
EMBL-EBI	European Bioinformatics Institute , part of the European Molecular Biology Laboratory
EU	European Union
HEP	High Energy Physics
ID	Identifier
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 <a href="http://orcid.org">http://orcid.org</a>
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	Publishing Network for Geoscientific & Environmental Data
PID	Persistent Identifier
PLOS	Public Library Of Science
PMC	PubMed Central
SIS	CERN Scientific Information Service
UI	Unique Identifier



## Appendix B: THOR 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<sup>23</sup>).

For more information, visit <http://project-thor.eu> or contact [info@project-thor.eu](mailto:info@project-thor.eu)

---

<sup>23</sup> <https://www.force11.org/group/joint-declaration-data-citation-principles-final>