Project Acronym: THOR
Grant Agreement no: H2020-EINFRA-2014-2 654039
Project Title: Technical and Human Infrastructure for Open Research

D2.1: Artefact, Contributor, and Organisation Relationship Data Schema

Document Information

Revision: Draft
Deliverable Type: Report
Dissemination Level: PUBLIC
Date: 11/09/2015

Authors: Martin Fenner (DataCite), Tom Demeranville (ORCID EU), Rachael Kotarski (BL), Todd Vision (UNC), Laura Rueda (DataCite), Robin Dasler (CERN), Laure Haak (ORCID), Patricia Cruse (DataCite)

Abstract: This document identifies gaps in existing PID infrastructures, with a focus on ORCID and DataCite Metadata and links between contributors, organizations and artefacts. What prevents us from establishing interoperability and overcoming barriers between PID platforms for contributors, artefacts and organisations, and research solutions for federated attribution, claiming, publishing and direct data access? It goes on to propose strategies to overcome these gaps.

DOI: http://doi.org/10.5281/zenodo.30799

Coordinated by the British Library, the THOR project is co-funded by the European Union.
Visit http://project-thor.eu for more information.

Revision history

Version	Status	Name, organisation	Date	Changes
0.1	Draft	Martin Fenner, DataCite Tom Demeranville, ORCID EU Rachael Kotarski, BL Todd Vision, UNC Laura Rueda, CERN Robin Dasler, CERN Laure Haak, ORCID EU	1-Sep-2015	
0.2	Draft	Martin Fenner, DataCite Patricia Cruse, DataCite Laure Haak, ORCID EU Tom Demeranville, ORCID EU Todd Vision, UNC	6-Sep-2015	Input from authors
0.3	Draft	Martin Fenner, DataCite	9-Sep-2015	Input from reviewers
0.4	Draft	Martin Fenner, DataCite	10-Sep-2015	Input from coordinator
1.0	Final	Adam Farquhar, BL	11-Sep-2015	Layout, minor changes

Review and approval

Action	Name, organisation	Date
Reviewed by	Jo McEntyre, EMBL-EBI Amir Aryani, ANDS Jamus Collier, PANGAEA	9-Sep-2015
Approved by	Adam Farquhar, BL	11-Sep-2015

Report template version: 1.3, 10-Sep-2015

Project summary

The **THOR** project will establish seamless integration between articles, data, and researchers across the research lifecycle. This will create a wealth of open resources and foster a sustainable international e-infrastructure. The result will be reduced duplication, economies of scale, richer research services, and opportunities for innovation.

The project has four concrete aims:

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

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

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

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

Copyright notice

Contents

1 Introduction

Great progress has been made in the implementation of persistent identifiers for contributors as well as for datasets and other artefacts, in particular in the last five years. There is a common understanding that persistent identifiers are foundational for all scholarly infrastructure. Work still needs to be done to implement support for these persistent identifiers in all relevant services, but the basic support infrastructure is already in place. In addition, uptake of these identifiers has increased dramatically since the start of the ODIN project in September 2012, with now for example more than 1.5 million ORCID identifiers and 6 million DataCite DOI names registered.

The next step to consider after the implementation of a basic support infrastructure for persistent identifiers for contributors and datasets is well underway, with a focus on enabling services that take advantage of this persistent identifier infrastructure. The THOR project will approach this from several angles, from building services (WP3), outreach activities to encourage the use and integration of PID services (WP4), to work on sustainability of this service infrastructure (WP5).

THOR WP2 deals with another important aspect of building this service infrastructure: **establish interoperability** and overcome barriers between PID platforms for contributors, artefacts and organisations, and **research** solutions for federated attribution, claiming, publishing and direct data access. As first steps in this work we identified the major gaps in the existing PID infrastructure, and proposed strategies to overcome these gaps. This effort is summarized in this document.

In our work we have focused on describing how the different PID platforms handle relations between persistent identifiers, in particular contributors to artefacts, contributors to organizations, and organizations to artefacts. We have taken two complementary approaches:

1. describe how ORCID and DataCite handle metadata for contributors, artefacts and organizations
2. describe PID workflows in two data centers (Archaeology Data Service and Dryad Digital Repository)

In both approaches we have included community standards where appropriate, and we have developed a detailed comparison table for all relevant metadata standards that is available in **Appendix A**. In the conclusions we not only discuss the major gaps that we have identified, but also propose further work in the following areas:

1. Common Approach to Personal Names
2. Standardized Contributor Roles
3. Standardized Relation Types
4. Metadata for Organisations
5. Persistent Identifiers for Projects
6. Harmonization of ORCID and DataCite Metadata

In the Appendix we not only list the relevant metadata vocabularies from ORCID, DataCite and the community, but also describe an example use of persistent identifiers for projects, using DataCite metadata and contributors and outputs from the ODIN project.

2 Existing Artefact, Contributor, and Organisation Relationship Data Schema

2.1 ORCID

2.1.1 Organisation - Contributor

The mission of ORCID is to maintain a registry of identifiers for researchers and contributors. It does not maintain a separate database of organisation identifiers, but rather leverages organisation identifiers provided by external providers such as ISNI and FundRef.

ORCID considers organisation-to-person connections a critical aspect of its functionality. ORCID's design specifications include persistent identifiers for organisations. ORCID follows best practice guidelines from NISO and uses Ringgold-ISNI identifiers for employment and education affiliation data.[1] Also following community guidance, ORCID uses FundRef identifiers to connect to funding information in user records.[2,3] The ORCID database structure accommodates multiple organisation identifiers for the same entity, acknowledging that more than one registry may be necessary to cover the organisations of interest to the community.[4]

ORCID ensures that each of its member organizations is associated with a Ringgold ID.[5] This ID is a component of how ORCID manages information provenance in its registry. Member universities that post affiliation information to employee ORCID records use Ringgold IDs and standard names; and the assertion of affiliation is tied back to the ID as well. The UberWizard tool that allows researchers to connect their ORCID record with funding data also uses organisation identifiers, in the form of FundRef identifiers.[6]

While ORCID strongly encourages the use of persistent identifiers for organisations, there are some instances where these identifiers are not mandatory. Some organisations may not be listed in the Ringgold database used to support ORCID's type-ahead manual affiliation functionality. In this case, ORCID allows the user to manually enter a name, and later works with Ringgold to consolidate and assign identifiers to new names. Organisation names added by an individual can be modified, but not names added through the API. In both cases, however, there is an identifier in the background. This means that different users may have a different description, or even name for an organisation, but that the record could point to the same uniquely identified organisation.

[1] See policy description in Haak, LL. 2013. ORCID Plans to Launch Affiliation Module using ISNI and Ringgold Organization Identifiers. ORCID Blog. http://orcid.org/blog/2013/06/27/orcid-plans-launch-affiliation-module-using-isni-and-ringgold-organization.

[2] See policy description in Haak, LL 2014. Link Your ORCID Record to Your Funding. ORCID Blog. http://orcid.org/blog/2014/02/19/link-your-orcid-record-your-funding.

[3] See xml for funding documentation: http://members.orcid.org/api/xml-funding.

[4] See xml for affiliations documentation: http://members.orcid.org/api/xml-affiliations.

[5] See Member Support Center documentation: http://members.orcid.org/api/organizations-orcid-ringgold-identifiers.

[6]http://www.uberresearch.com/uberwizard-for-orcid-launched-supporting-researchers-in-adding-grants-from-various-funders-to-their-orcid-records-with-a-free-and-open-tool/

Funding, employment, and education connections can be seen as 'top level' relationship types, each having multiple sub-types from defined vocabularies. In all cases, ORCID encourages users to query the originating source for authoritative metadata.

2.1.1.1 Funding

ORCID considers funding to be an artefact in its own right, with its own persistent "funder" identifier, derived from the FundRef Registry.[7] Funding can be "grant", "contract", "award" or "salary-award". In common with most other ORCID activities, such as journal articles, funding activities include metadata about contributors. ORCID is currently working with others in the community to implement a standard contributor taxonomy.[8]

This metadata is ideally provided by funders following the grant award, in which case the funder can validate the information and connection with the ORCID record. Funding connections may also be entered by the record holder manually or using a search and link wizard currently supported by UberResearch. The contributor list may contain several ORCID identifiers, referencing the record holder and other persons, if that information was collected by the research funder at time of grant submission.

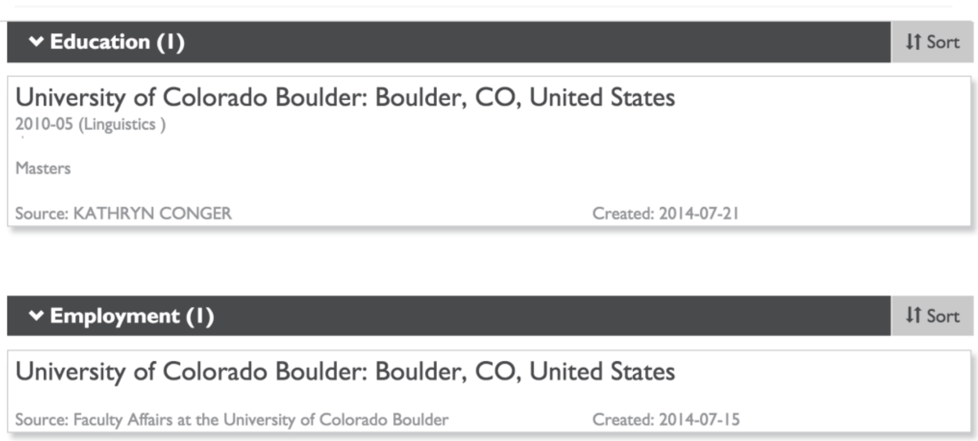
2.1.1.2 Employment

Employment activities are linked with ORCID records via an identifier and organisation name, and can include information on location (state/country), department, title, and start and end date. Users may enter this data manually, taking advantage of a type-ahead prompt for organisation name that will auto-associate the entry with an identifier and location information. In this case, the record will show that the source of the information was the record holder. It is also possible for employers that are ORCID members to post affiliation information into an ORCID record, with the record holder's permission. This process has been used by a number of universities, and is described in the Create and Connect documentation.[9] In this case, the source will show the member-employer's name, and that member will have the ability edit the record, such as when the person leaves the organisation. Figure 1 shows an example of a record that shows affiliation and education information with the individual and the organisation listed at the sources.

[7] http://www.crossref.org/fundref/fundref_registry.html.

[8] See Paglione L. 2015. Contributor Role Update. ORCID Blog. http://orcid.org/blog/2015/08/11/contributor-recognition-update-orcid-project-credit-and-contributorship-badges.

[9] http://members.orcid.org/create-records.



Education (1)

Sort

University of Colorado Boulder: Boulder, CO, United States
2010-05 (Linguistics)

Masters

Source: KATHRYN CONGER

Created: 2014-07-21

Employment (1)

Sort

University of Colorado Boulder: Boulder, CO, United States

Source: Faculty Affairs at the University of Colorado Boulder

Created: 2014-07-15

Figure 1. Example of ORCID record affiliation and education information, derived from different sources. From http://orcid.org/0000-0002-9949-4025.

2.1.1.3 Education

Education activities are represented in the same way as employment. Member organisations are using the Create and Connect workflow to connect students, their thesis, and the organisation in an authenticated pathway that yields a validated public electronic record of a dissertation award. This is valuable to an individual during their career, and to the degree-granting institution, which can use the ORCID APIs to easily track the individual’s career contributions.

2.1.2 Artefact - Contributor

Artefacts are commonly known as ‘activities’ in the ORCID ecosystem. Activities encompass all research outputs as well as the previously mentioned funding, employment and education. Research outputs are collectively known as ‘works’ and can have one of the 39 ORCID work types listed in **Appendix B**.

Works have their own, limited, set of metadata attached to them, and ORCID mandates that works data added through the API include at least one persistent unique identifier. This metadata is a subset of that used by the original source, mapped to the ORCID schema[10]. In addition to DOIs, 30 other identifier schemes are supported. They also contain contributor metadata, in the same way as funding records. ORCID contributor roles are listed in **Appendix B**. ORCID has also started to incorporate the contributor roles from Project CRediT[11] (**Appendix D**) into their Registry.[12]

Works have an optional ‘citation’ field. This can contain a formatted citation in a recognised format such as Harvard or APA, or it can contain a BibTeX-formatted citation. However, lack of consistency and machine-readability makes the contents of this field hard to leverage.

Works have ‘sources’. This captures who was responsible for pushing artefact metadata into an ORCID record and can include the ORCID record holder. This is a provenance relationship. It should be noted that record holders may connect works information to their record using search and link wizards, in

[10] The ORCID metadata schema was determined in consultation with the community. See the report by the Metadata Working Group. http://members.orcid.org/api/supported-work-identifiers

[11] http://credit.casrai.org/

[12] http://orcid.org/blog/2015/08/11/contributor-recognition-update-orcid-project-credit-and-contributorship-badges

which case the works metadata will likely be accurate. In addition to self-claim by record holders, employers may also push works data to ORCID records. Different sources may have different levels of assurance or 'trust' and it is up to the user of the data to decide which source to trust. ORCID displays multiple connections to works independently, so it is possible to see all sources for a given artefact for a given profile via the user interface and the API, making it straightforward to check assertions for any item in the record.

2.1.3 Artefact - Organisation

ORCID allows records only for individuals. Organisations cannot register for an ORCID identifier. Relationships between individuals and organisation can be obtained via an API query for an organisation identifier, which will return information on associated ORCID identifiers and identifiers for works, etc. For example, it would be possible to determine which grant numbers are associated with which organisations, although that would currently require a full scan of ORCID records.

2.2 DataCite

2.2.1 Organisation - Contributor

Within DataCite, contributors and organisations do not exist as separate entities. Rather, they are attached to one or more DOI names as part of the *creator* and/or *contributor* attributes. Consequently, a relationship between organisations and contributors can't be described directly with DataCite metadata, but only indirectly via linking both to one or more DOI names. With this limitation DataCite can't describe a relationship between organisations and contributors unless there is a work linking them, and then the relationship is indirect.

In contrast to ORCID, organisations can be creators or contributors of a work, and some contributor roles (e.g. *HostingInstitution*) are specific to organisations. In addition, creators and contributors can have an affiliation, which is most appropriate if the creator or contributor is an individual. Affiliations are described as text strings and there is no specific field for persistent identifiers – whereas when an organization is a creator or contributor we could use the *NameIdentifier* field.

2.2.2 Artefact - Contributor

DataCite uses two attributes to describe an artefact - contributor relationship: *creator* and *contributor*. Both attributes can be used for individuals and organisations, and the metadata does not indicate whether the *creator* or *contributor* is one or the other. *Creator* is a required metadata field and includes, in priority order, *the main researchers involved in producing the data, or the authors of the publication*. *Contributors* are *the institution or person responsible for collecting, managing, distributing, or otherwise contributing to the development of the resource*.

DataCite uses a two-tiered approach to describe the work type: a *ResourceTypeGeneral*, which uses a controlled vocabulary, and a *ResourceType*, which is free text. Both are currently optional fields, although ResourceTypeGeneral is planned to be a mandatory field in the next major release of the

DataCite schema. *ResourceTypeGeneral* was modelled after the DublinCore resource type list[13] and is listed in **Appendix E**.

Both *creators* and *contributors* have the attributes *name*, *nameIdentifier*, and *affiliation*. In addition *contributors* can also have the attribute *contributorType*. In other words, creators are a subgroup of all contributors; they always have the contributorType *creator,* and they are a required attribute. The contributorType uses a controlled vocabulary, the list of possible values is provided in **Appendix F**. Some of these contributorTypes are more appropriate for individuals, whereas other roles are more appropriate for organisations.

The name field for *creators* and *contributors* are creatorName and contributorName, respectively. In both cases the name is entered into a single field, rather than separate fields for given and family names as in the case of ORCID. For personal names the format should be *family, given*. The nameIdentifier attribute supports multiple name identifier schemes through the nameIdentifierScheme attributes. The affiliation attribute is a free-text field and there is no special field for an affiliation persistent identifier.

2.2.3 Artefact - Organisation

Artefact - organisation relationships are not handled any differently from artefact - contributor relationships, as described above where organisations can be creators or contributors. Organisations can also be included via the *affiliation* field of a creator or contributor. Since this is a text field rather than a special field for persistent identifiers, it is hard to build artefact - organisation relationships via the *affiliation* field.

2.3 Gaps

When trying to align the ORCID and DataCite metadata schemata, the following gaps were identified:

1. different work type vocabularies between DataCite and ORCID
2. different contributor role vocabularies between DataCite and ORCID
3. different approaches to personal names, one field vs. two fields plus name variants
4. authorship by consortia/projects - supported by DataCite, but not by ORCID
5. different approaches to link funders: to the work in DataCite vs. to the contributor in ORCID. A deeper problem is that artifacts can be funded by multiple funders via multiple people and these relationships are lost both in the ORCID and DataCite metadata

In **Appendix A** we compare the ORCID and DataCite metadata schemata in detail via a table, and also look at some other common vocabularies, including Dublin Core, CASRAI and MODS. We discuss the proposed steps to close these gaps at the end of this document.

3 Prototype Persistent Identifier Submission Workflows

The previous section looked at how persistent identifiers and other metadata are handled in the ORCID Registry and DataCite Metadata Store. In this section we want to look at the transfer of information between scholarly infrastructures, starting with two example data centers. To identify the gaps in this

[13] http://dublincore.org/documents/resource-typelist/

use case, we have looked at the way in which data archives are currently exchanging information on artefacts, contributors and organisations (using identifiers) with DataCite, ORCID, but also with publishers of linked articles.

As part of the ORCID and DataCite Interoperability Network (ODIN) project, a generic workflow for ORCID and DataCite identifiers was developed that illustrated how data archives can include and manage these two classes of identifiers throughout the data archiving lifecycle.[14]

Here we have overlaid the processes of two data archives onto the generic workflow: The Archaeology Data Service (ADS), based in York, UK and Dryad, an international repository for data underlying publications in science and medicine. To expand on the generic ODIN workflow, we have highlighted the actual steps these repositories take in managing identifiers, and how these processes integrate with artefacts held outside of the repository – specifically articles.

3.1 The Archaeology Data Service (ADS)

Archaeological research is by its nature destructive. Once a site has been excavated, it cannot be re-examined in its original context. The ADS was set up in 1996 to provide an archive for the outputs of archaeological research from a wide variety of sectors. The ADS holds data produced by third sector researchers in organisations such as local archaeological trusts, private organisations providing pre-construction archeological investigation, local government and government departments, as well as academic research.

This wide variety of contributors means they have differing objectives for supplying research objects to the ADS and so differing needs in terms of identifying themselves and other contributors to that work. Some of these were highlighted in the ODIN output.[15]

The ADS workflow contributed to the development of the ODIN generic workflow, but concentrated on individuals as the contributors and did not look at the integration with anyone other than DataCite and ORCID. Figure 1 shows the ADS workflow overlaid on the generic workflow - with matching processes highlighted and unused areas of the generic workflow greyed out.

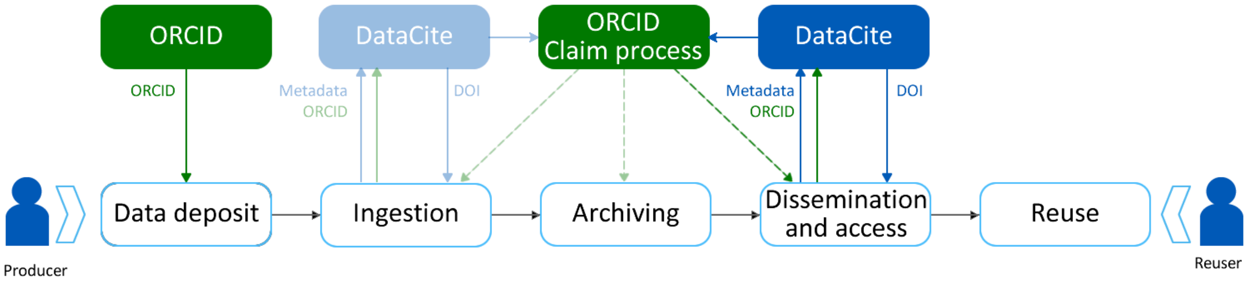


Figure 1. A workflow diagram showing identifier flow through the repository archiving process for the ADS and Dryad. Highlighted areas are those where the repository process matches with the generic workflow developed in ODIN.

[14] Josh Brown, Amir Aryani, Amy J Barton, Jan Brase, Tom Demeranville, Patricia Herterich, et al. (2015). D4.2: Workflow for interoperability. Figshare. http://doi.org/10.6084/M9.FIGSHARE.1373669

[15] Rueda L, Dallmeier-Tiessen S, Kotarski R, Newbold E, Herterich P, Lavasa A, Brown Josh (2015) ODIN D3.3: Proofs of concept and commonality. Figshare http://dx.doi.org/10.6084/m9.figshare.1373665

The ADS has a close relationship with the open access journal *Internet Archaeology*.[16] The journal is produced within the same department as the ADS and *Internet Archaeology*'s publishing process uses the ADS's collection management system.

Internet Archaeology use the 'People' table within the ADS's database to store information on authors. This allows immediate 'exchange' of identifiers between the archive and the journal. Although such a close infrastructural tie is unlikely to occur in many other archive-journal relationships, the information exchanged and the stages at which it occurs may still be appropriate in other contexts.

3.1.1 Contributor Identifiers

The ADS would like to get name identifiers for contributors at deposit, although in reality, this happens rarely at present. *Internet Archaeology* request ORCID identifiers on submission, although may not receive them until later in the manuscript process, if authors have had to create their ORCID identifiers later on. In either case, the ADS is either able to 'quietly' update its metadata for records related to authors who supply their ORCID identifiers to *Internet Archaeology*, or use the ORCID identifier if the author subsequently submits data to the ADS.

For older materials, the ADS make ad-hoc appeals to their contributors to inform the ADS of their ORCID identifiers. The ADS then associate ORCID identifiers with contributors, which updates all records associated with those contributors. This additional metadata can then be passed on to DataCite in bulk-updates of metadata. This process requires effort on the part of the ADS in appealing for the information, entering it into their content management system and then updating DataCite. Integrating ORCID claims directly from ORCID back into their system would reduce that effort.

3.1.2 Organisation identifiers

Contributor and creator records within the ADS are associated with organisations. Organisations may also be the creators and contributors themselves. These organisations are currently given the ADS internal identifiers, but there is no reason they could not be associated using organisational identifiers via FundRef and/or ISNI. If these identifiers could be supplementary on ORCID claims, it would reduce the effort required for the ADS to make use of external identifiers.

The roles given to persons within the ADS are also used interchangeably as roles for organisations. Some of the roles given to organisations and persons within the ADS are used for the ADS' administrative purposes and so may not be useful outside of ADS. The ADS does supply information on roles to DataCite via the DataCite metadata, but only for items it classes as Archives rather than individual reports[17].

[16] *Internet Archaeology* (205). http://intarch.ac.uk/

[17] 'Archives' in ADS are collections of research objects associated with a specific project (excavation, survey, scientific analysis, etc.) or piece of work. Reports are distinct, individual items that document a particular investigation.

3.1.3 Content Identifiers

The ADS uses DataCite DOIs as persistent identifiers for all of its research objects. *Internet Archaeology* assigns CrossRef DOIs to all its articles. As the journal and archive are using a shared collection management system, these identifiers are shared between the archive and journal as soon as they are created.

3.1.4 Integrating archiving and publishing workflows

In terms of where the ADS' archiving workflow fits into the publishing workflow of *Internet Archaeology* or vice versa, either submission of data to the ADS or submission of an article to *Internet Archaeology* could be the starting point. As they share infrastructure, both parties can closely follow the development of the other as items go through the process whether submissions occur in parallel or sequentially. The result for the current situation is that they get the ORCID identifiers of new submissions from each other once (whether given at data or manuscript submission), and can fetch any other information they need from ORCID.

This also results in each party being able to access the ORCID identifiers for submissions at a number of points in the workflow. Figure 2 gives a guide to the current points of exchange of identifiers between the ADS and *Internet Archaeology*. Where ORCID identifiers are exchanged, organisational identifiers could potentially also be exchanged. As these workflows can run sequentially or in parallel, the points of exchange of identifiers for people and organisations can run in both directions. They can also flow between multiple parts of each workflow – dashed lines indicate one of multiple points where the identifier could be exchanged.

In terms of retrieving information on claims via ORCID, a claim of a single item in ORCID that the ADS could verify, would result in all other records from that contributor in the ADS (as well as in *Internet Archaeology*) also being associated with that ORCID identifier, as there is only a single point of update required within the ADS's collection management system. The position of the ORCID Claim process in Figure 2 is shifted in comparison to the ODIN generic workflow. This is to clarify that the claims are generated during the Reuse phase of the data and in our two examples for this work, those claims only then add metadata back into the dissemination package.

With parallel submission and processing of the article and data, the identifiers for both are shared at later points in the workflow. Although they can share information on what the DOI will be before publication, this may not allow them to retrieve metadata from DataCite or CrossRef – the DOI name can be known internally, but may not be minted until publication. *Internet Archaeology* manually adds the DOI name for items in the ADS to their articles where appropriate.

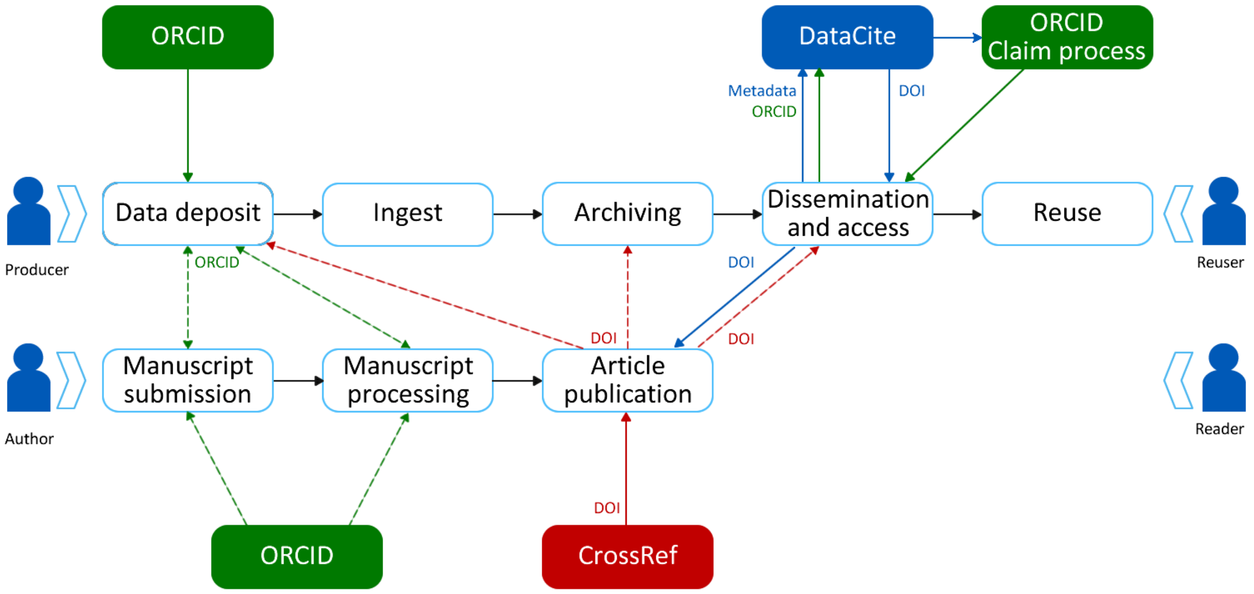


Figure 2. Exchange of identifiers between repository and journal as part of the integrated publishing and archiving workflows of both the ADS and Dryad. The ‘Producer’ to ‘Reuser’ process is that of the data repository, and the ‘Author’ to ‘Reader’ process shows that of the journal.

3.2 Dryad Digital Repository

The Dryad Digital Repository hosts research data underlying findings in scientific and medical publications, particularly ‘long-tail’ data for which dedicated domain repositories are lacking. A ‘data package’ in Dryad is defined as the set of files associated with a single publication together with the metadata describing the contents. The associated publication is most commonly a journal article, but may also be a monograph, dissertation or other scholarly work. Dryad has workflows in place whereby publishers may integrate the submission of data with the submission of manuscripts, so that metadata can be exchanged between the repository and publisher prior to publication (similar to the exchange between ADS and *Internet Archaeology*) and data may be privately accessed by editors and peer reviewers prior to manuscript acceptance.

3.2.1 Contributor identifiers

Dryad is adopting ORCID identifiers for a number of functions: as a trusted identifier that can be shared in public metadata, as a source of information for clustering and disambiguating author names, and as a preferred login mechanism for depositors and their collaborators. Currently, ORCID identifiers are linked to identities in the system by a call from the Dryad submission system to the ORCID API that enables a depositor to look up and select their ORCID identifier and those of their coauthors. ORCID identifiers are not required metadata for each name field. Dryad will not require an ORCID identifier for each name in the system, but will instead build enhanced services on top of ORCID identifiers to motivate their uptake.

As ORCID identifiers become more widely used, it is anticipated that Dryad will be able to rely more frequently on two other mechanisms: (1) receiving ORCID identifiers in the metadata provided by manuscript submission systems of integrated publishers, and (2) harvesting ORCID claims for Dryad-associated publications from ORCID profiles themselves (and possibly other sources such as CrossRef), after publication. Having author-name-to-ORCID mappings for Dryad-associated publications will go most, but not all of the way, toward having complete author-name-to-ORCID coverage within Dryad. The list of contributors to the data package is, by default, the same as the list of authors on the corresponding publication. However, submitters have the option to add, remove, and rearrange the order of contributors to the data package as a whole (which is the container level at which Dryad data are cited) and may similarly edit contributor lists for the component data files.

As with the ADS, Dryad would update DataCite metadata with ORCID identifiers if they are discovered after the initial DOI registration.

3.2.2 Organisation identifiers

Dryad does not currently use organisation identifiers but has several features in development that will require them. First, in order to better support implementation of funder data policies, Dryad is moving towards capturing funding information as an optional field. It is anticipated that FundRef will provide close-to-complete coverage of the funding sources for Dryad's contributors, but this remains to be determined in practice. As with ORCID identifiers, FundRef information can be captured pre- or post-publication, and either directly from contributors or indirectly from manuscript submission systems. Funder metadata at the organisational level can be conveyed to DataCite as part of initial DOI registration or via updates. A current gap is that individual project and grant numbers are not yet modelled within DataCite's metadata schema; harmonizing the treatment of individual project and grant information between CrossRef and FundRef would be desirable for repositories, such as Dryad, that work with content funded by many different organisations around the globe.

Another intended use of organisational identifiers is for the institutions with which contributors are affiliated. Such information would primarily be useful for institutions tracking, and in some cases paying for, the data outputs of their researchers. Dryad plans to test the suitability of ISNIs as identifiers for organisational affiliations as a first step. As with the prior use of organisation identifiers, this information may be captured pre- or post-publication and either directly from contributors or indirectly from manuscript submission systems. It is expected that affiliation metadata would be conveyed to DataCite as part of initial registration or via updates.

Persistent identifiers would also be useful to Dryad to identify non-individual contributors. Such contributors are becoming increasingly common in the life sciences. However, many of the non-individual contributors that receive authorship credit are, in fact, grant-funded projects of limited duration (e.g. The 1000 Genomes Project Consortium[18]) rather than formal organizations, and so unlikely to be included within current institutional registries such as ISNI and FundRef. Thus, such organisations fall between the cracks of our current identifier infrastructure and cannot currently be supported. Persistent identifiers for projects (see below) are one possible approach.

[18] 1000 Genomes Project Consortium et al. (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491, 56-65. http://10.1038/nature11632

3.2.3 Content identifiers

Dryad mints DataCite DOIs for data packages as well for the component data files. The data package DOI is shared with the publisher so that it can be included in the publication, and some publishers take responsibility for that step. For other publishers, and for all non-integrated submissions to Dryad, the responsibility for assuring the data package DOI is included within the publication rests with the author. A relatively small number of submissions to Dryad are for data associated with existing publications, in which case assuring a link from the publication back to the data is more challenging (although still possible through services such as CrossMark[19], Elsevier's data linking service[20] or Europe PMC BioEntity links[21]). The identifier for the publication (almost always a CrossRef DOI) is a required field in the data package metadata, and so Dryad generally assures that a link from the data package to publication will be present. However, for those data packages released before publication, there may be a lag before the publication identifier is known.

3.2.4 Integrating archiving and publishing workflows

Much of the identifier metadata will be mirrored between Dryad and a publisher (e.g. identifiers authors, affiliations, funders, DOIs) and so standardized communication with the publisher prior to publication is the primary way Dryad efficiently assures metadata correctness and comprehensiveness when a data package is released. Dryad is integrated with over 70 journals from publishers that use a variety of manuscript submission systems (including Editorial Manager, eJournal Press, ScholarOne, and Open Journal Systems). Publishers choose, for each journal, at which of three different timepoints in the publishing workflow data submission occurs: (i) before the author submits the manuscript, (ii) after manuscript submission and before review, or (iii) after review and acceptance. Different identifiers may be known at different stages of the process, with typically author identifiers (and their affiliations) known the earliest and CrossRef and DataCite DOIs the latest. Because of this staggered provision of identifiers, metadata exchange between the publisher and repository can, for some integrated journals, be a multistep process. Nonetheless, the general ODIN model can be applied to the case of Dryad, as depicted in Fig 2. For the approximately one quarter of Dryad's data packages that are associated with non-integrated publications, the identifiers must either be provided manually by the author during submission or harvested (e.g. from CrossRef and ORCID) after publication.

3.3 Gaps

Despite their differences, the ADS and Dryad both support a close relationship between data archiving and publication, and thus there are many similarities in workflow that may not be shared with other data repositories, in particular the multistage exchange of identifiers directly between the publisher and repository prior to publication and data release.

This analysis highlights a number of gaps in identifier infrastructure that remain to be addressed. Doing so would be of considerable value to both publication and data archiving workflows.

1. Taking greater advantage of ORCID claims: Currently both the ADS and Dryad appeal for creators to register for ORCID identifiers and enter them into records, but uptake is patchy and there

[19] http://www.crossref.org/crossmark/
[20] http://www.elsevier.com/connect/bringing-data-to-life-with-data-linking
[21] http://europepmc.org/help#bioentitieslinks

remains considerable legacy content. Importing claims information from the ORCID Registry would improve coverage, even if it would sometimes occur post-publication, and even if claims were more often available for publications than data. It would also enable repositories to update DataCite metadata with ORCID identifiers, which neither the authors nor ORCID are authorized to do directly.

2. Coverage of organisational affiliation within ORCID: Organisational affiliations currently require manual entry for each author of each work. If organisational identifiers were more reliably available within ORCID profiles, this would make the submission workflow more efficient for publishers, repositories and contributors, as well as allow institutions to better track their outputs.
3. Organisations as contributors and creators: Organisational contributors are needed in the ADS for items from commercial and government sources, and in Dryad for items from large collaborative projects. However, many of these organisations do not currently have identifiers from ISNI or FundRef. What can be done to ensure these contributors do have organisational identifiers? Guidelines for assigning appropriate roles to organisations within DataCite metadata would also be useful.
4. Granularity of funding information: Grant and project information is currently supported by FundRef, but the DataCite metadata schema only includes this information to the level of the funding organisation, which makes it difficult for funders and institutions to, for instance, track compliance with individual data management plans.

4 Conclusions and Future Work

In this work we described the current status of linking datasets and other artefacts to contributors and organisations (both funders and institutions). We looked at the ORCID and DataCite metadata, at other community standards, and at the practical implementation in two data centers.

We found that the workflows of linking contributors to datasets are in place, but that there are a number of gaps that still need to be addressed. When we looked at additional information linked to these contributors and artefacts – in particular funding information and contributor roles – we found that this information is not yet handled in a consistent way and further work is needed.

4.1 Common Approach to Personal Names

DataCite and ORCID use a different approach to contributor names: DataCite uses a single field for the name, whereas ORCID splits the name into two fields. Both approaches are common in other scholarly systems that capture personal names,[22] including data centers, and there is no common approach prevalent in the scholarly community. When using a single field to capture names we see both *family name, given names* (the DataCite recommendation), and *given names family name*. Going from a single field to two fields is more error-prone than the other direction, and two fields are sometimes needed (e.g. sorting by family name, abbreviation of given names to initials).

[22] http://www.w3.org/International/questions/qa-personal-names

Separate fields for given and family names are required for proper formatting of citations.[23] As long as citations to scholarly content rely on properly formatted text rather than persistent identifiers, services holding bibliographic information have to support these separate fields. We therefore recommend that all services collecting bibliographic information use two separate fields for personal names. We also need a third field for contributors that are organizations, as this information shouldn't be put into personal fields to avoid confusion when joining name parts together.

To facilitate the transition of services using single input fields for names, e.g. DataCite, the THOR project will generate documentation about best practices for handling personal names based on existing guidelines[24,25], and will either list tools that help in this transition, e.g. the namae[26] parser for the Ruby language, or will help generate these tools where needed. As a first step THOR will collect information about existing practices in the various communities served by THOR partners.

4.2 Standardized Contributor Roles

As the list of contributors to a particular research output such as a dataset is constantly increasing, and research outputs become increasingly diverse, encompassing not only text, but also data, software, workflows, and more, we need to better understand the role of each individual contributor in producing the research output. This is particularly important for datasets and other outputs that currently don't receive the same amount of attention compared to journal articles.

The existing contributor role vocabularies provided by ORCID (**Appendix C**) and DataCite (**Appendix F**), not only have little overlap, but also describe the general role (e.g. data manager) rather than the individual contribution. In the case of DataCite, contributor roles are not applicable to creators, the main contributors to a work.

Project CRediT (**Appendix D**) is a multi-stakeholder initiative that has developed a common vocabulary with 14 different contributor roles, and this vocabulary can fill this gap of describing the specific contribution of a contributor. CRediT is complementary to existing contributor role vocabularies such as ORCID conributor roles and DataCite contributor types. For contributor roles it is particularly important that the same vocabulary is used across stakeholders, so that for example the roles assigned to the creators of a dataset in a data center can be forwarded first to DataCite, then to ORCID, and then also to other places such as institutional repositories.

The implementation of a common contributor role vocabulary is a major challenge. The THOR partners will discuss how best to approach this in future work in WP2, based on implementation work that ORCID is already doing, and based on community input.

4.3 Standardized Relation Types

Similar to contributor roles, different vocabularies are in use for relation types between persistent identifiers. Rather than describing the relation between a contributor (or sometimes organization) and an artefact, relation types typically describe the type of relation that exist between two artefacts, e.g. A

[23] http://docs.citationstyles.org/en/stable/specification.html#names
[24] http://www.ncbi.nlm.nih.gov/books/NBK7282/
[25] https://readthedocs.org/projects/citation-style-language/
[26] https://github.com/berkmancenter/namae

is a new version of B, or A references B. As relations often exist between persistent identifiers issued by different persistent identifier services, we need a common vocabulary to describe these relations. ORCID is not tracking relations between works (other than multiple versions of the same work contributed from different sources), but DataCite has a controlled vocabulary of relation types, listed in **Appendix H**. This list is based on the Dublin Core relations,[27] but has been extended over time. Other organizations use different vocabularies to describe relation types, e.g. CrossRef.[28]

Capturing relations in a standardized way is important, as they are needed for citations and thus the basis for many indicators of scholarly impact. Some scholars reserve the term citation for links between two scholarly articles, and the term data citation is used both for all links from scholarly works to datasets, or reserved for formal citations appearing in reference lists.

Relation types are sometimes described through the work types of the link items, e.g. a Wikipedia page referencing a scholarly article.

One challenge of relations between artefacts is that they are temporal in nature, i.e. work A might cite work B years after both work A and B have been published. This means that any work on standardizing relations has to include not only services registering persistent identifiers for works, but also services tracking links between artefacts. THOR will do further research and community work to suggest a common vocabulary of relation types that works across artefacts. This includes work on collecting links to artefacts provided by THOR partners, and describing the most common relation types found.

4.4 Metadata for Organisations

Both ORCID and DataCite not only provide persistent identifiers for people and data, but they also collect metadata around these persistent identifiers, in particular links to other identifiers. The use of persistent identifiers for organisations lags behind the use of persistent identifiers for research outputs and people. Despite the work by ISNI, FundRef and others, community uptake is still low. In addition, for some of these organizational identifiers (e.g. FundRef) there is no openly available central service that systematically collects links to other identifiers. Funders and institutions have their own internal systems to track these links (e.g. to people and data), but there is no common approach that would for example allow a third party to find all datasets produced at a particular institutions, or all people funded by a particular funder. Some of these questions can be answered by querying the ORCID Registry or DataCite Metadata Store, but these queries are time-consuming and the information is often incomplete.

The THOR partner institutions need to work on a common approach to better expose and aggregate links to organisations. This includes:

1. promote the use of persistent identifiers for organizations where appropriate. In particular, support the use of FundRef identifiers for funders. Europe PMC and other THOR partners will develop best practice guidelines for this implementation
2. Track uptake of persistent identifiers by THOR partners in the WP5 metrics dashboard
3. enable the ability to use persistent identifiers for organizations where this is currently not possible. Ideally, use actionable identifiers for organizations

[27] http://dublincore.org/documents/usageguide/elements.shtml#relation
[28] http://www.crossref.org/help/schema_doc/4.3.5/relations_xsd.html

4. make sure that persistent identifiers for organisations are propagated from data centers to DataCite, from institutions to ORCID, and from DataCite to ORCID
5. develop strategies to aggregate information about organisations available in the ORCID Registry and DataCite Metadata Store. This could be done by working with third parties, or by ORCID and/or DataCite building additional services
6. work with CrossRef to also include research outputs with DataCite DOIs in FundRef Search[29]

4.5 Persistent Identifiers for Projects

Research projects are collaborative activities among contributors that may change over time. Projects have a start and end date and are often funded by a grant. The existing persistent identifier infrastructure does support artefacts, contributors and organisations, but there is no first-class PID support for projects. This creates a major gap that becomes obvious when we try to describe the relationships between funders, contributors and research outputs. Both the ORCID and DataCite metadata support funding information, but only as direct links to contributors or research outputs, respectively. This not only makes it difficult to exchange funding information between DataCite and ORCID, but also fails to adequately model the sometimes complex relationships, e.g. when multiple funders and grants were involved in paying for a research output. Furthermore, in certain disciplines projects are cited as “authors”, something that existing PID infrastructure fails to model.

This gap can only be addressed by using persistent identifiers for projects, with the following characteristics:

1. open infrastructure that does not create any barriers for entry based on geographic region, discipline or cost
2. ideally based on existing persistent identifier infrastructure to decrease the cost and duration of implementation
3. ability to add metadata, in particular links to contributors, organizations and artefacts, and central registry to search these metadata

Grants are the most important use case for project identifiers, and funders should therefore be deeply involved in the discussion about persistent identifiers for projects. One persistent identifier that fits the above criteria is the DOI, and as an example use case we have created the DOI http://doi.org/10.5281/zenodo.30030 for the EC-funded ODIN project (the predecessor to THOR), and linked all contributors and research outputs (see **Appendix I**). The most important use cases are already covered with this DOI, and relatively little work would be needed to adapt the DataCite Metadata Schema to fully support projects (adding *ResourceTypeGeneral* Project, and adding *DateTypes* for project start and end dates). Ideally funders all use their own specific DOI prefix, and funders could of course use a DOI registration agency different from DataCite, e.g. Publications for Europe in the case of EC funding. Using DOIs is one of several possible implementation strategies for project identifiers. THOR will work with all stakeholders to make progress in the area of persistent identifiers for projects by raising awareness for this important gap, building community support and piloting technical implementations.

[29] http://search.crossref.org/fundref

4.6 Harmonization of ORCID and DataCite Metadata

We identified significant differences between the two metadata schemata, and these differences hinder the flow of information between the two services. Several different approaches to overcome these differences are conceivable:

1. only use a common subset, relying on linked persistent identifiers to get the full metadata
2. harmonize the ORCID and DataCite metadata schemata
3. common API exchange formats for metadata

The first approach is the linked open data approach, and was designed specifically for scenarios like this. One limitation is that it requires persistent identifiers for all relevant attributes (e.g. for every creator/contributor in the DataCite metadata). One major objective for THOR is therefore to increase the use of persistent identifiers, both by THOR partners, and by the community at large. Increased uptake of persistent identifiers that enable cross-linking is more important than collecting detailed information in every service separately. THOR will use a combined strategy of further research, service development and community outreach to achieve that goal. One consequence of this approach is that individual persistent identifier services, e.g. ORCID or DataCite, will hold a version of record associated with the persistent identifier - e.g. how to write a particular name or the exact title of a work - rather than the still common practice of having metadata maintained in multiple places, resulting in differences that are sometimes difficult to reconcile.

A good use case is the detailed information about contributors that ORCID is collecting (name variants, education, employment, etc.). There is no need for DataCite to collect this information as well, as it can link to ORCID via the ORCID name. The THOR partners will discuss with the community what metadata need to be made available across all services (for instance, those needed for a citation) compared to information that can be held in specific registries such as ORCID or DataCite and queried on demand or imported when linking identifiers.

A common metadata schema between the two organisations is neither feasible (due to different use cases; different governance, etc.) nor necessarily desired. In addition, we have to also consider interoperability with other metadata standards (e.g. CASRAI, OpenAIRE, COAR) and other artefacts, such as those having CrossRef DOIs. What is more realistic is harmonization on a smaller scale, for example using a common format for essential metadata, which should minimally include

1. **for contributors**: family name, given names, persistent identifier and identifierScheme (could be multiple)
2. **for artefacts**: title, persistent identifier and identifierScheme, contributors (with fields as in 1.), publication date
3. **for organizations**: name, persistent identifier and identifierScheme

THOR WP2 will work on refining this list of essential metadata, e.g. whether work type, affiliation or publisher should also be essential metadata shared across services. Consistent implementation of these metadata is also critical, not only for personal names (see above), but also for dates (how to handle partial dates, etc.) and persistent identifiers (expressed as namespace plus ID or as URL, etc.).

The other area besides essential metadata where harmonization between services is required is describing the relations between contributors and artefacts, or between artefacts and other artefacts. We need a common vocabulary if we want to describe details beyond the link between two persistent identifiers, e.g. to describe contributor roles or relation types. We therefore propose work on **standardized contributor roles** and **standardized relation types** (see previous sections).

For other metadata this might often not be feasible, and we can use a translation table to translate for example work types from the DataCite schema to the ORCID schema (**Appendix G**), using the mapping between DataCite and ORCID[30] as example. Citeproc JSON[31] is well-suited as a common exchange format because mappings of work types already exist,[32] and the format is already used by several DOI registration agencies including CrossRef and DataCite,[33] as well as ORCID.[34]

The third approach to improve interoperability uses a common API format that includes all the metadata that need to be exchanged, but doesn't require the metadata schema itself to change. This approach was taken by DataCite and CrossRef a few years ago[35] to provide metadata for DOIs in a consistent way despite significant differences in the CrossRef and DataCite metadata. Using HTTP content negotiation, metadata are provided in a variety of formats[36]. A pilot for ORCID HTTP content negotiation was started in the ODIN project[37], and work is underway to implement this functionality in the ORCID production service. Citeproc JSON (see also previous paragraph) is of particular interest, as this machine-readable format is used by CrossRef, DataCite and ORCID, as well as many other services (in particular reference managers).

Ultimately the harmonization of ORCID and DataCite metadata will depend on a combination of the three approaches mentioned above, serving different communities and use cases. Having identified the major gaps, we can start closing those gaps in upcoming work in WP2 and WP3, in particular **D3.3** (Services that enable integration and cross-linking across different types of identifiers and data types).

[30] https://github.com/crosscite/doi-metadata-search/blob/master/lib/datacite/work_type.rb
[31] http://gsl-nagoya-u.net/http/pub/citeproc-doc.html
[32] http://gsl-nagoya-u.net/http/pub/csl-fields/index.html
[33] http://crosscite.org/cn/
[34] http://feed.labs.orcid-eu.org/
[35] http://crosstech.crossref.org/2012/05/crossref_and_datacite_unify_su.html
[36] http://crosscite.org/cn/
[37] http://feed.labs.orcid-eu.org/

Appendix A. Comparison Table Contributors, Artefacts and Organizations

The following table provides a concise comparison of the metadata across schema.

CONTRIBUTORS

DataCite	ORCID	Dublin Core	Dublin Core terms namespace	CASRAI dictionary	MODS	DDI Lifecycle 3.2	Comments
<creator>	<orcid-bio>	dc:creator	dcterms:creator				
<creatorName>	<given-names>	dc:creator	dcterms:creator	Person Info/Salutation	<name>		While <creatorName> in DataCite is repeatable, with each
<nameIdentifier>	<external-identifiers>	dc:identifier	dcterms:identifier	Person ID Types	<identifier>	<a:ResearcherIdentification>	
<nameIdentifier	<external-identifier>			(name identifier scheme is	<identifier type=" ">	<a:TypeOfID>	
<contributor>	<work-contributors>	dc:contributor	dcterms:contributor	Research Dataset Contributor		<r:Contributor>	ORCID adds contributors to the works
<contributorType>	<contributor>	dc:contributor	dcterms:contributor	Research Dataset Contributor/Role	<name><role>	<r:ContributorRole>	ORCID <contributor-role> of "author" maps to DataCite's
<contributorName>	<contributor>	dc:contributor	dcterms:contributor	Research Dataset	<name>	<r:ContributorName>	
<nameIdentifier	<contributor-orcid>			Research Dataset Contributor/ID	<identifier>	<a:TypeOfID>	ORCID only supports ORCID iDs for contributors

ARTEFACTS

DataCite	ORCID	Dublin Core	Dublin Core terms namespace	CASRAI dictionary	MODS	DDI Lifecycle 3.2	Comments
<title>	<work-title>	dc:title	dcterms:title	*/[Object]Title	<titleInfo>	<r:Title>	
<title titleType=" ">	<work-title>	dc:title	dcterms:alternative	Work type is a parent element to	<titleInfo>	<r:AlternateTitle>	DataCite has a controlled list of values for titleType:
	<journal-title>			Journal Article/Journal	<relatedItem type="host"> +		
<publisher>		dc:publisher	dcterms:publisher	*/[Object]Publisher	<originInfo><publisher>	<r:Publisher>	
<publicationYear>	<publication-date>	dc:date	dcterms:issued	*/Publication Date	<originInfo><dateIssued>	<r:SimpleDate>	DataCite only supports year, ORCID supports month. Dublin
<subjects>		dc:subject	dcterms:subject		<subject authority=" "> (for controlled		DataCite allows free text. Dublin Core Subject element also
<dates>		dc:date	dcterms:date	(other dates as sub-elements of	<oiriginInfo>	<r:SimpleDate>	DataCite has a controlled list of values for dateType: accepted,
<language>	<language-code>	dc:language	dcterms:language		<language><languageTerm>	<r:Language>	DataCite and ORCID both accept ISO 639-1 (and DataCite also
<resourceType	<work-type>	dc:type	dcterms:type	Output Types list	<typeOfResource> (controlled)	(defers to Dublin Core)	ORCID uses the CASRAI Output Standard.
<alternateIdentifiers> <alternateIdentifier alternateIdentifierType=" ">	<work-external-identifiers> <work-external-identifier> <work-external-identifier-type> <work-external-identifier-id>	dc:identifier	dcterms:identifier	Output ID Types list	<identifier type=" ">	<r:UserID typeOfUserID=""> <a:CallNumber>	CASRAI Data Identifier term can be a DOI, Handle, etc. Digital Object Identifier term is reserved for DOIs. Neither term is a sub-element of the other.
<relatedIdentifiers> <relatedIdentifier relatedIdentifierType=" " relationType=" " relatedMetadataScheme=" " schemeType=" " schemeURI=" ">		dc:relation + dc:identifier dc:relation (for relationType)	dcterms:relation + dcterms:identifier Dublin Core relation types: dcterms:conformsTo dcterms:isReferencedBy dcterms:references dcterms:isVersionOf dcterms:hasVersion dcterms:isFormatOf dcterms:hasFormat dcterms:isPartOf dcterms:hasPart dcterms:isRepalcedBy dcterms:replaces dcterms:source		<relatedItem> + <identifier>	<r:UserID> <r:UserID typeOfUserID=""> (for relatedIdentifierType) <r:Relationship> <r:RelationshipDescription>	Can also map to Source element in Dublin Core, depending on the relationship specified by DataCite.
<size>		dc:format	dcterms:extent		<physicalDescription><extent>	<r:Content> <pi:CaseQuantity> (if unit "cases" is known) <a:DataFileQuantity> (if unit "datafile" is known)	DataCite accepts free text
<format>		dc:format	dcterms:format		<physicalDescription><form> <physicalDescription><internetMediaType >	<pd:FileFormat> <a:ItemFormat>	DataCite accepts free text, MIME if posible. Dublin Core recommends controlled MIME type for Format element.
<version>						<pi:PhysicalInstance version="">	Suggested practice: track major_version.minor_version
<rightsList> <rights rightsURI=" ">		dc:rights	dcterms:rights		<accessCondition>	<a:Access> <a:AccessConditions>	DataCite accepts free text
<description>	<short-description>	dc:description	dcterms:description		<abstract>	<r:Description>	
<geoLocations> <geoLocation> <geoLocationPoint> <geoLocationBox> <geoLocationPlace>				Research Location/Location Geo Tag	<subject><cartographic><coordinates> (for point and/or box coordinates) <subject><hierarchicalGeographic> (for machine parsable geographic names) <subject><geographic> (for geographic subject terms that aren't parsed)	<r:GeographicLocation>	
	<work-citation> <work-citation-type> <citation>		dcterms:bibliographicCitation			<r:Citation>	

ORGANISATIONS							
DataCite	ORCID	Dublin Core elements namespace	Dublin Core terms namespace	CASRAI dictionary	MODS	DDI Lifecycle 3.2	Comments
<affiliation>	<affiliations> <affiliation> <type> <department-name> <role-title> <start-date> <end-date> <organization> <name> <address> <disambiguated-organization> <disambiguated-organization-identifier> <disambiguation-source>			HE Employment History/Institution	<name><affiliation>		DataCite supports <affiliation> both for <creator> and <contributor> and allows free text

See http://doi.org/10.5281/zenodo.30799 for the underlying data.

Appendix B. ORCID Work Types[38]

1. artistic-performance
2. book-chapter
3. book-review
4. book
5. conference-abstract
6. conference-paper
7. conference-poster
8. data-set
9. dictionary-entry
10. disclosure
11. dissertation
12. edited-book
13. encyclopedia-entry
14. invention
15. journal-article
16. journal-issue
17. lecture-speech
18. license
19. magazine-article
20. manual
21. newsletter-article
22. newspaper-article
23. online-resource
24. other
25. patent
26. registered-copyright
27. report
28. research-technique
29. research-tool
30. spin-off-company
31. standards-and-policy
32. supervised-student-publication
33. technical-standard
34. test
35. translation
36. trademark
37. website
38. working-paper

[38] See Documentation at http://members.orcid.org/api/supported-work-types.

Appendix C. ORCID Contributor Roles

1. author
2. assignee
3. editor
4. chair-or-translator
5. co-investigator
6. co-inventor
7. graduate-student
8. other-inventor
9. principal-investigator
10. postdoctoral-researcher
11. support-staff
12. Lead
13. “Co lead”
14. “Supported by”

Appendix D. Contributor Roles from Project CRediT[39]

1. **Conceptualization**. Ideas; formulation or evolution of overarching research goals and aims.
2. **Methodology**. Development or design of methodology; creation of models.
3. **Software**. Programming, software development; designing computer programs; implementation of the computer code and supporting algorithms; testing of existing code components.
4. **Validation**. Verification, whether as a part of the activity or separate, of the overall replication/reproducibility of results/experiments and other research outputs.
5. **Formal analysis**. Application of statistical, mathematical, computational, or other formal techniques to analyse or synthesize study data.
6. **Investigation**. Conducting a research and investigation process, specifically performing the experiments, or data/evidence collection.
7. **Resources**. Provision of study materials, reagents, materials, patients, laboratory samples, animals, instrumentation, computing resources, or other analysis tools.
8. **Data curation**. Management activities to annotate (produce metadata), scrub data and maintain research data (including software code, where it is necessary for interpreting the data itself) for initial use and later re-use.
9. **Writing – original draft**. Preparation, creation and/or presentation of the published work, specifically writing the initial draft (including substantive translation).
10. **Writing – review & editing**. Preparation, creation and/or presentation of the published work by those from the original research group, specifically critical review, commentary or revision – including pre- or post-publication stages.
11. **Visualization**. Preparation, creation and/or presentation of the published work, specifically visualization/data presentation.

[39] http://credit.casrai.org/proposed-taxonomy/

12. **Supervision**. Oversight and leadership responsibility for the research activity planning and execution, including mentorship external to the core team.
13. **Project administration**. Management and coordination responsibility for the research activity planning and execution.
14. **Funding acquisition**. Acquisition of the financial support for the project leading to this publication.

Appendix E. DataCite ResourceTypeGeneral from Metadata Schema 3.1[40]

In parentheses are the number of active DOI names with that ResourceTypeGeneral as of September 4, 2015. 2,436,261 (39.8%) of all active DOI names have no ResourceTypeGeneral (an optional attribute)

1. Audiovisual (1,625)
2. Collection (220,333)
3. Dataset (1,705,103)
4. Event (501)
5. Image (462,877)
6. InteractiveResource (217)
7. Model (311)
8. PhysicalObject (181)
9. Service (18)
10. Software (7,156)
11. Sound (112)
12. Text (443,286)
13. Workflow (18)
14. Other (848,474)

[40] http://doi.org/10.5438/0011

Appendix F. DataCite ContributorType from the Metadata Schema 3.1

In parentheses are the number of active DOI names with that ContributorType (an optional attribute) as of September 4, 2015.

1. ContactPerson (522,206)
2. DataCollector (17,608)
3. DataCurator (3,169)
4. DataManager (282,241)
5. Distributor (2,309)
6. Editor (62,315)
7. Funder (12,615)
8. HostingInstitution (1.034,060)
9. Producer (5,960)
10. ProjectLeader (6,265)
11. ProjectManager (8)
12. ProjectMember (124)
13. RegistrationAgency (171)
14. RegistrationAuthority (1)
15. RelatedPerson (5,118)
16. Researcher (215,763)
17. ResearchGroup (3,845)
18. RightsHolder (24,945)
19. Sponsor (847)
20. Supervisor (6,743)
21. WorkPackageLeader (18)
22. Other (55,781)

Appendix G. Mapping of Work Types from ORCID and DataCite to Citeproc

ORCID	DataCite	Citeproc
book		book
book-chapter		chapter
book-review		review-book
dictionary-entry		entry-dictionary
dissertation		thesis
encyclopedia-entry		entry-encyclopedia
edited-book		
journal-article		article-journal
journal-issue		

ORCID	DataCite	Citeproc
magazine article		article-magazine
manual		
online-resource	InteractiveResource	webpage
newsletter-article		article-newspaper
report	Text	report
research-tool		
supervised-student-publication		
test		
translation		
website		webpage
working-paper		article
conference-abstract		paper-conference
conference-paper		paper-conference
conference-poster		paper-conference
disclosure		
license		
artistic-performance		
data-set	Dataset	dataset
invention		
lecture-speech		speech
research-technique		
spin-off-company		
standards-and-policy		
technical-standard		
	AudioVisual	motion_picture
	Image	Graphic
	Sound	Song

Appendix H. DataCite RelationType from the Metadata Schema 3.1

IsCitedBy
Cites
IsSupplementTo
IsSupplementedBy
IsContinuedBy
Continues
HasMetadata
IsMetadataFor

IsNewVersionOf
IsPreviousVersionOf
IsPartOf
HasPart
IsReferencedBy
References
IsDocumentedBy
Documents
IsCompiledBy
Compiles
IsVariantFormOf
IsOriginalFormOf
IsIdenticalTo
IsReviewedBy
Reviews
IsDerivedFrom
IsSourceOf

Appendix I. Prototype Project encoded as DOI

Using the EC-funded ODIN project (the predecessor to THOR) as an example, we created a DOI for the ODIN project with the public summary information of the project attached as PDF and with the following information.

DOI
http://doi.org/10.5281/zenodo.30030

Title
ORCID and DATACITE Interoperability Network (ODIN)

PublicationYear
2012

ResourceTypeGeneral
Other

Description
‘Data as infrastructure’ is a critical concept for a fully-integrated European Research Area (ERA) to drive innovation forward as envisaged by the Digital Agenda for Europe. The lack of data availability hinders this vision. In academic publishing, peer review and citation have long been recognised as mechanisms for endorsing the trustworthiness of research outputs and incentivizing researchers to contribute. Trustworthy research data will only be widely available if the same principles are applied. Key, participative, initiatives have emerged to address this challenge.

The DataCite consortium has assigned over 1m DOI names in the last few years to make research data citable, true to emergence of the ‘4th paradigm’, Jim Gray’s vision of “data-intensive scientific discovery”.

ORCID offers the opportunity to identify individual authors across systems and infrastructures, including scholarly works that can have up to thousands of authors (as in the case of the LHC project). Researchers often change their affiliation, and collaborate across national disciplinary boundaries.

ODIN aims to build on the success of DataCite and ORCID by designing an 'awareness layer' for persistent author and object identifiers, thereby reducing technical, cultural and logistical barriers to the accessibility, attribution and trust of data. Identifier awareness will make it possible to stabilise: References to a data object; Tracking of use and re-use; Links between a data object, subsets, articles, rights statements and every person involved in its life-cycle (creator, editor, reviewer, aggregator, etc.).

Given the importance of these functions as we approach HORIZON2020, we aim to prove the feasibility of author, data and rights identification, promote trust building towards open scientific data e-Infrastructures and lay the foundation necessary to promote future interoperability (technical, semantic, reference architecture, etc.) in the scientific data domain in Europe and globally.

Contributor
European Commission (Funder, info:eu-repo/grantAgreement/EC/FP7/312788)

Creators
Aryani, Amir (ANDS)
Brase, Jan (DataCite)
Brown, Josh (ORCID EU)
Burton, Adrian (ANDS)
Dallmeier-Thiessen, Sünje (CERN)
Demeranville, Tom (British Library)
England, Jude (British Library)
Fenner, Martin (ORCID EU)
Herterich, Patricia (CERN)
Haak, Laurel (ORCID EU)
Kotarski, Rachael (British Library)
Lavasa, Artemis (CERN)
Mele, Salvatore (CERN)
Rueda, Laura (CERN)
Ruiz, Sergio (DataCite)
Thorisson, Gudmundur (ORCID EU)
Vision, Todd (Dryad)
Warner, Simeon (Cornell University)
Ziedorn, Frauke (DataCite)

Related Works
The research outputs below were all at least in part funded by the ODIN grant. All relations are of relationType IsPartOf.

10.1504/IJKL.2014.069537
10.5281/zenodo.21429
10.5281/zenodo.21430
10.6084/m9.figshare.1373671
10.6084/m9.figshare.1373670
10.6084/m9.figshare.1373669
10.6084/m9.figshare.1373665

10.6084/m9.figshare.1373668
10.6084/m9.figshare.824316
10.6084/m9.figshare.1373664
10.6084/m9.figshare.843603
10.6084/m9.figshare.825546
10.6084/m9.figshare.824314
10.6084/m9.figshare.824315
10.6084/m9.figshare.824317
10.6084/m9.figshare.107019
10.6084/m9.figshare.154691
10.6084/m9.figshare.1373666
10.6084/m9.figshare.824318
10.6084/m9.figshare.154690
10.6084/M9.FIGSHARE.1057958
10.5281/ZENODO.10521

The following information could also be added to the DataCite metadata, but this was not possible via the Zenodo upload interface:

ResourceType
Project

Contributor (HostingInstitution)
ANDS (ISNI, 0000 0004 4663 7787)
British Library (ISNI, 0000 0001 2301 1923)
CERN (ISNI, 0000 0000 9547 8293)
Cornell University (ISNI, 0000 0004 1936 877X)
DataCite (no ISNI)
Dryad (ISNI, 0000 0004 4663 8050)
ORCID (ISNI, 0000 0004 4663 8501)

Although the DataCite metadata can contain dates of various types, the *DateType* controlled vocabulary doesn’t include adequate descriptions for project start and end dates (2012-09-01 and 2014-09-30 in this case). One possibility with the existing metadata would to be to use *Date Valid* with the date range 2012-09-01 to 2014-09-30.