THOR: Conceptual Model of Persistent Identifier Linking

Document Information

Date: 31/03/2016

Authors: Martin Fenner (DataCite), Tom Demeranville (ORCID EU), Rachael Kotarski (BL), Robin Dasler (CERN), Johanna McEntyre (EMBL-EBI), Guilherme de Mello (EMBL-EBI), Todd Vision (DRYAD), Angela Dappert (BL), Adam Farquhar (BL)

Reviewer: Markus Stocker (PANGAEA)

Abstract: In this report we describe the current state of the art for persistent identifier linking in scholarly e-Infrastructure, with a focus on persistent identifiers for contributors and data. We look at persistent identifier linking between datasets, for example different versions of the same data, as well as linking data with other resources, including articles, contributors, institutions, and funding information.

DOI 10.5281/zenodo.48705

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 1 March 2016.

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

Contents

Executive Summary: Conceptual Model of Persistent Identifier Linking 1

1 Introduction 1

2 Existing Persistent Identifier Linking Implementations by THOR Partners 3

2.1 ORCID 3

2.1.1 Multiple Records 3

2.1.2 Versioning 3

2.1.3 Organisational Identifiers 4

2.2 DataCite 5

2.2.1 Alternate and Related Identifiers 5

2.2.2 Versioning 6

2.2.3 Organisational Identifiers 6

2.3 EMBL-EBI 7

2.4 CERN 8

2.4.1 Alternate Author Identifiers 8

2.4.2 Identifier Linking and Versioning 10

2.4.3 Organisational Identifiers 11

2.5 British Library 11

2.5.1 Versioning 11

2.5.2 ISNI as Institutional Identifier 12

2.6 PANGAEA 13

2.6.1 Identifier Linking 13

2.6.2 Versioning 14

2.6.3 Relation Types 14

3 Conceptual Model for Persistent Identifier Linking 15

3.1 Linkage as Triples 15

3.2 Describing the Relation 15

3.3 Persistent Identifiers as HTTP URIs 16

3.4 Centralised Infrastructure for Identifier Linking 17

3.5 Further Work 17

4 Discussion 19

4.1 Identifier Linking for Versioning Data 19

4.1.1 Major vs. Minor Versions 20

4.1.2 Landing Pages 20

4.1.3 Future Work: Notifications 21

4.1.4 Future Work: Dynamic Datasets 21

4.2 Identifier Linking for Data and Articles 21

4.3 Data Linking with Contributors, Institutions and Funders 22

4.3.1 Contributors 22

4.3.2 Institutions 23

4.3.3 Funding Information 23

5 Conclusions 24

6 References 25

Appendix A: Project summary 27

Appendix B: Terminology 28

Executive Summary: Conceptual Model of Persistent Identifier Linking

In this report we describe the current state of the art for persistent identifier linking in scholarly e-Infrastructure, with a focus on persistent identifiers for contributors and data. We look at persistent identifier linking between datasets, for example different versions of the same data, as well as linking data with other resources, including articles, contributors, institutions, and funding information. We describe common practices, including those in data repositories from four different disciplines, and identify shortcomings in the existing implementations. Based on this survey of current practice, we propose a conceptual model for persistent identifier linking that can aid implementation and adaptation, with a focus on scalability. Finally, we describe three important use cases for persistent identifier linking: versioning of data, linking articles and data, and linking data with contributors, institutions and funding information. We highlight areas where development of services is taking place or anticipated, and identify challenges that need further work.

1 Introduction

Persistent identifiers (PIDs) provide long-lasting, globally unique references to digital and physical objects. They are a core component of scholarly e-Infrastructure, help build the scholarly record, and allow us to discover truth by building on previous discoveries, by *standing on the shoulders of giants*[1]. The EC-funded THOR project[2], via its partners ORCID and DataCite, is providing persistent identifiers for contributors and data. THOR is working with partners at the British Library, CERN, PANGAEA, EMBL-EBI, ANDS, Dryad, Elsevier and PLOS on services linking these persistent identifiers, improving workflows, and providing training.

Although the term *persistent identifier* is commonly used, these identifiers have other characteristics beyond persistence. A 2013 report from the ODIN project (ODIN Consortium et al., 2013) calls them *trusted identifiers* and describes them as digital identifiers which are *unique*, *persistent*, *descriptive*, *interoperable* and *governed*. The Joint Declaration of Data Citation Principles (Martone, 2014) recommends a persistent method for identification of datasets that is machine actionable, globally unique, and widely used by a community. The 2011 Den Haag Manifesto (Knowledge Exchange, 2011) recommends that PIDs should be HTTP URIs that support content negotiation and provide access to a minimum common set of metadata elements across different kinds of identifiers used in scholarly communication.

A common theme of these recommendations is that persistent identifiers need to be machine actionable. In practical terms, this means that they can be referred to as persistent HTTP URIs. Machine actionable linking is different from linking for humans. The latter requires context and enables discovery by an individual human user, but doesn't enable automatic linking between resources on a larger scale.

[1] https://en.wikipedia.org/wiki/Standing_on_the_shoulders_of_giants

[2] http://cordis.europa.eu/project/rcn/194927_en.html

Robust persistent identifier infrastructure provides:

- **Specificity**: linking identifiers to track different versions and different levels of granularity of a research output
- **Evidence**: linking identifiers when citing other scholarly works or resources
- **Credit and attribution**: linking identifiers for research outputs to identifiers for contributors, institutions and funders

This report looks at how persistent identifiers can best support linking between resources, with a focus on identifier linking for contributors and data.

In Section 2, we start by describing the situation at two key persistent identifier service providers: ORCID[3] and DataCite[4]. We move on to explore scenarios in four different stakeholders: major research institutions in the life sciences and physics (EMBL-EBI, CERN); a key data centre in earth and environmental sciences (Pangaea), and a national research library (The British Library). We describe current practices and commonalities in their approach to persistent identifier linking, but also identify open issues. One important observation of this work is that persistent identifier linking is done in different ways, depending on the kind of resource that is linked, the persistent identifier used, and the organisation doing the linking. For both end-users and service providers, this variety requires extra effort. Creating and discovering links between scholarly resources is more complex and more costly than it needs to be.

In Section 3, we build on this insight to introduce a conceptual model for identifier linking that simplifies some of the complex variation across systems and thus improves scalability of services providing these linkages.

In Section 4, we describe how this conceptual model can help with identifier linking to provide specificity, evidence, and credit and attribution.

Throughout, we focus on three important use cases:

- Identifier linking for specificity – linking for versioning of data
- Identifier linking for evidence – linking data and articles
- Identifier linking for credit and attribution – looking at linking of data with contributors, institutions and funding information

The findings presented in this report will inform future work of the THOR partners in implementing and improving services for persistent identifier linking, help other e-Infrastructure providers implement persistent identifier linking, and better scale services.

[3] http://orcid.org

[4] http://datacite.org

2 Existing Persistent Identifier Linking Implementations by THOR Partners

2.1 ORCID

ORCID is an organisation that creates and maintains a registry of unique persistent researcher identifiers. It provides a transparent method of linking research activities concerning education, employment, and funding, as well as research outputs to these identifiers[5].

2.1.1 Multiple Records

Although multiple records for the same person are not common within the ORCID system, ORCID does not mandate that a person only has a single ORCID record. There are cases when multiple records are desirable. Situations where an author does not want to attach certain outputs to their main scholarly record and would prefer to use a pseudonym exist in many disciplines. In these cases the records are not linked within ORCID: they appear to the outside world as two distinct people. ORCID does provide a set of policies and practices to prevent unintentional duplicates being created, and mechanisms for resolving these into single records.

To prevent the creation of unintentional duplicates, upon creation ORCID suggests existing accounts based on name and email addresses, and alerts users to the fact that they may already have a record. This is complemented by a forgotten password functionality and the ability for ORCID accounts to have multiple email addresses associated with them.

To resolve existing duplicates, the ORCID support team is able to deprecate one account and have it point to another, which is then marked as the primary. ORCID relies on individuals reporting duplicates and requires proof of ownership and authorisation to deprecate/merge records. A dispute procedure exists for situations where more than one individual is involved and a complaint is filed. Deprecation is preferred over deletion to maintain the persistence of the ORCID iD. Deprecated records viewed via the web or API contain a link to the primary record and nothing else.

2.1.2 Versioning

ORCID maintains relationships between the ORCID iD and work identifiers such as DOIs. The ORCID registry is deliberately unaware of the particular versioning strategies employed by the creators of those identifiers; it is up to the ORCID record owner to decide which links are created and maintained. This version-agnostic approach can be seen as a complement to versioning strategies applied by data centres. The data centre can maintain the full complement of versions and version metadata in the manner most suitable to their datasets. The user is then free to decide which is the most relevant version to attach to their scholarly record, and at what granularity it should be included.

ORCID does maintain multiple associations between an ORCID record and the identifier for a specific work. Multiple associations arise when more than one source (for example a publisher) asserts that the ORCID record owner contributed to a uniquely identified work. For example, a user could manually

[5] http://orcid.org

enter a DOI that is subsequently also added by the publisher or institutional repository. Works with multiple sources are grouped together within the user interface and API, with the user specifying the preferred source. This source and the metadata associated with it are then displayed on the web and at the top of the API representation.

Multiple sources that add the same identifier can assert different metadata within the ORCID record, such as an alternative title or publisher. A common example for journal articles is metadata from CrossRef vs. PubMed. This approach enables the owner to have complete control over what is exposed via the web and API. However, it should be noted that the identifier remains the same: a DOI will resolve to the authoritative source no matter how the ORCID record describes it. In some cases, identifiers cannot be relied upon to uniquely identify a work. For example, an ISBN identifies a whole book where in many cases the ORCID record owner contributed a single chapter. In these cases, ORCID does not group works by identifier and leaves it to the user to maintain a single source per distinct work.

The decision on whether to group is based on the identifier type. If a DOI were to resolve to a compound item with complex authorship such as a book or dataset, then ORCID would be ignorant of this and group the identifier regardless. There is currently no way to 'ungroup' multiple sources, and any attempt to do so would be complex from both a user interface and technical point of view.

2.1.3 Organisational Identifiers

ORCID uses two external identifiers to model relationships between organisations and activities: Ringgold[6] and CrossRef Funding Data (formerly FundRef)[7]. Ringgold provides unique organisation identifiers for use by subscribers. It is a registration agency for ISNI, the International Standard Name Identifier for people and organisations[8]. CrossRef Funding Data is a registry of identifiers for funding organisations. It is available via an open license.

Activities that can be linked to organisations include employment, education and funding. CrossRef Open Funder Registry IDs are used to identify organisations involved in funding activities, such as a research council; Ringgold identifiers are used to reference organisations involved in funding, employment and education activities, such as a university. The use of organisation identifiers for activities is optional. The ORCID user interface uses the controlled list from Ringgold to auto-suggest organisation names to the user when creating funding, employment or education activities for their record. These suggestions can be ignored, allowing users to manually enter organisations that are not within the Ringgold system. Users cannot enter external identifiers themselves.

Funding sources require a combination of organisation identifier and grant number to be uniquely identified. As there is no single source of grant identifiers, this is a free-form field that does not reference to any system outside of the funding organisation. Typically, a funding activity contains the grant number and a URL, which resolves to a human readable page about the grant hosted by the awarding body. External clients such as ResearchFish[9], CRIS[10] systems or ÜberWizard for ORCID[11] can

[6] http://www.ringgold.com/

[7] http://www.crossref.org/fundingdata/

[8] http://isni.org

[9] https://www.researchfish.com/node/2525

choose between Ringgold, CrossRef Open Funder Registry ID, or opt not to include an identifier when adding funding sources to ORCID records.

There are, however, limitations. Alternative spellings, although captured by Ringgold and exposed through the user interface, are not exhaustive. Alternative languages present the same problem. Departmental, project or discipline allegiances may trump institutional ones, for example providing 'Project X' over 'University Y', which results in the ORCID registry having many ORCID records that are not associated with an external organisation identifier. Instead, reference organisation identifiers are linked through employment or education activities. Consequently, individual users may reference the same entity in different ways, and ORCID has no way of automatically identifying these matches.

ORCID holds metadata that maps Ringgold to ISNI on a one-to-one basis, although this is not yet exposed through the user interface or API.

2.2 DataCite

DataCite is a leading global membership organisation offering reliable persistent data identification services. Its mission is to make research better by enabling people to find, share, use and cite data. DataCite is a registration agency for DOI names. As of early 2016, it works with around thirty members around the globe, and supports over 600 data centres to assign DOIs to research data and other scholarly outputs.

2.2.1 Alternate and Related Identifiers

All resources that have a DataCite DOI as their persistent identifier can have additional unique identifiers associated with them in the DataCite metadata. This is assigned using the optional **alternateIdentifier** field, which is paired with **alternateIdentifierType** – a field for describing the identifier used. Both are free-text fields. As of February 12, 2016, 1,384,529 records use the **alternateIdentifier** field (query), so it is already in common use. Given the free-text format of both fields, it is difficult to systematically extract information from these fields without further work on database indexing.

Related Identifiers are identifiers for other resources linked to the DataCite record via the **relatedIdentifier** field. 1,639,030 records use the **relatedIdentifier** field (query) as of February 12, 2016. The relation type is defined via a controlled vocabulary (for example "IsNewVersionOf", "IsCitedBy"), as is the type of related identifier (for example "DOI", "Handle" or "URL"). An important use case for the **relatedIdentifier** field is the description of links between data and articles. Because of the relation type "IsIdenticalTo", there is overlap with the **alternateIdentifier** field: those alternate identifiers that use one of the allowed values for **relatedIdentifierType** can also be included as related identifiers.

[10] Current Research Information System, more info at http://www.eurocris.org/

[11] https://orcid.uberresearch.com/

2.2.2 Versioning

Version information is an optional field in the DataCite metadata (DataCite Metadata Working Group, 2014), using a free-text field. No specific formatting is required for the version field, but DataCite recommends “major_version.minor_version”, based on work by ESIP (Earth Science Information Partners)[12].

DataCite recommends:

- Issuing a new DOI for a major version change
- Linking the different versions of a dataset using the relation types “IsNewVersionOf” and “IsPreviousVersionOf”
- Describing the version in the **description** field

What constitutes a new version, and whether or not to issue a new DOI, is a decision made by the data centre, as practices vary across disciplines.

While version information is typically used to indicate changes over time, versions can also represent particular forms of a dataset, for example translations. This differs from varying representations of the same content, for example in different file formats, for which the DataCite metadata has a dedicated **format** field. Versioning information is included in multiple DataCite metadata fields (**identifier**, **version**, **description**, **relatedIdentifier**), the use of which differs across data centres. For example, not all data centres use a **relatedIdentifier** with relation type “IsNewVersionOf”. This inconsistency makes it difficult to collect detailed information about versioning across data centres.

As of February 12, 2016, the version information field is used in 1,611,740 DataCite metadata records (query), and the “IsNewVersionOf” relation type 262,209 times (query). Within the **resourceTypeGeneral** dataset, 659,329 out of 1,635,084 (40.3%) metadata records include version information. These numbers indicate that version information is seen as important by many data centres.

2.2.3 Organisational Identifiers

Three DataCite metadata elements refer to organisation information: **creator**, **contributor** and **affiliation**. Both people and organisations can be creators or contributors, and there is no flag to distinguish between the two. Contributors can be further characterised via the **contributorType** controlled vocabulary. Contributor types that apply to organisations include Funder, HostingInstitution, RegistrationAgency, ResearchGroup and Sponsor. HostingInstitution is a commonly used field, present in 942,385 records (query) as of February 12, 2016. The Funder attribute is only used in 13,453 records (query) as of February 12, 2016. Persistent identifiers for contributors, whether an individual or an organisation, can be included in the optional **nameIdentifier** field in conjunction with the **nameIdentifierScheme** field.

In the DataCite 4.0 Schema that will be released in 2016, funding information will be encoded differently, using a separate **FundingReference** field. This will make it easier to encode grant identifiers (via **AwardNumber** and **AwardURI**), and will include **FunderIdentifier** and **FunderIdentifierType** fields. These fields are consistent with how funding information is encoded in CrossRef metadata. These

[12] http://wiki.esipfed.org/index.php/Interagency_Data_Stewardship/Citations/provider_guidelines#Note_on_Ver%20sioning_and_Locators

changes in recording funding information in the DataCite metadata will hopefully increase the percentage of metadata records with funding information.

Affiliation is a free text field for creators and contributors, and was added in the 3.1 schema in 2014. Once the field is added to the DataCite Metadata Store index, we can explore how many DOI records use that field, and what values are most commonly used. In *Artefact, Contributor, and Organisation Relationship Data Schema* (Fenner et al., 2015), we argued that to achieve harmonisation of ORCID and DataCite Metadata, detailed information about contributors, including their affiliation, should be stored in the ORCID registry; DataCite would then link to the ORCID record via the nameIdentifier attribute for creators and contributors.

re3data[13] is a global registry of research data repositories from a wide range of disciplines. It provides additional information about each repository, including description, licence information, subject areas, and a unique identifier for each metadata record. In 2016, re3data became a DataCite service. Work is ongoing to link re3data identifiers with the DataCite internal identifiers for data repositories.

2.3 EMBL-EBI

The European Bioinformatics Institute[14] is part of the European Molecular Biology Laboratory. EMBL-EBI provides freely available data from life science experiments, performs basic research in computational biology and offers an extensive user training programme, supporting researchers in academia and industry (Cook, C. et al., 2015).

The EMBL-EBI hosts many data resources that support research in the life sciences[15]. Typically the resources are the result of global collaborations, many of which have been in operation for several decades. The databases are organised around different types of biological data, from nucleotide sequences to protein structures to chemicals. Resources are either archival databases, which take *de novo* data submissions from researchers, or added-value knowledge bases, which re-use and curate data to provide views on the data from a biological perspective. The latter save researchers significant time, organising and clustering information in understandable interfaces.

All data resources use identifier systems that are at least unique within the resource. These identifiers are widely known within their respective communities, and are used to refer to data in research articles and in database cross links. As an example, the European Nucleotide Archive (ENA) is a submission database, operating since the late 1980s. It ingests and shares nucleotide sequences into a cross-continental system, the INSDC (EMBL-EBI, nd), which also includes GenBank in the USA and the DDBJ in Japan. Over 700 million sequences have identifiers in formats familiar to the community that submits to and uses ENA[16]. Data submissions can come from individual scientists in experimental labs; yet increasingly they are automatically ingested from sequencing machines. All are archived with rich biological metadata

[13] http://www.re3data.org/

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

[15] http://www.ebi.ac.uk/services/all

[16] See: http://www.ebi.ac.uk/ena/about/statistics and http://www.ebi.ac.uk/ena/about/citing-ena-data

in standard formats that enable scientific searches and analysis. While ENA has many direct users, the added-value knowledge bases are one of its most prominent consumers. One of these is Ensembl, which provides genome assemblies computed from ENA records. Another is UniProt, a heavily used and expertly curated database of proteins, which is a hub of information for many data resources including ENA[17]. This pattern is repeated across many resources, building a rich network of data that would not be possible without robust identifier and versioning systems, protocols, and community standards. Figure 1 shows the cross-linking of major life sciences resources; the width of the connecting ribbons reflects the relative volumes of data shared.

Some resources can have multiple identifiers, often reflecting the evolution of a resource. For example, in Europe PMC, an article record can have a PMID, PMCID and a DOI. Protein Data Bank (PDB) has both PDB identifiers and DOIs, and a small number of other data resources at the EBI use DOIs as a secondary identifier alongside the operational one: ChEMBL, for example, assigns DOIs to its data releases, and the PRIDE database of proteomics assigns DOIs to 'full submissions'. Yet this practice is not widespread at the EMBL-EBI.

A further challenge is the fact that several databases operate across international consortiums. For example, within PDB, a given PDB identifier can resolve to a number of different landing pages – in fact, the DOI resolves directly to a data download at the PDB FTP site. The EBI operates a service called identifiers.org[18], which emerged from a requirement in systems biology to resolve biological objects in a system model to a resource unambiguously. It tracks life science database identifier patterns, locations that a given database-identifier pair may resolve to, and has the potential to extend to a wider role in cataloguing and organising life science data resources, their identifiers and their resolution.

2.4 CERN

The European Organization for Nuclear Research (CERN) operates the largest high energy physics (HEP) laboratory in the world. It uses extensive digital infrastructure to support this research, including INSPIRE[19], the High Energy Physics Information System.

2.4.1 Alternate Author Identifiers

When ORCID was first integrated in INSPIRE, ORCID iDs were manually added by Scientific Information Service (SIS) personnel. With the new INSPIRE Labs, researchers are now able to authenticate via ORCID to claim their author profile. If users authenticate with ORCID in INSPIRE, information is pushed to their ORCID profile. Manual ORCID addition is still possible, but it does not push the INSPIRE record to the ORCID profile. At this time, only publications are appended to the ORCID record; existing metadata found in the ORCID record is not corrected. Record corrections will be implemented in the future.

[17] See, for example: http://www.uniprot.org/uniprot/Q13077#sequences

[18] http://identifiers.org

[19] http://inspirehep.net

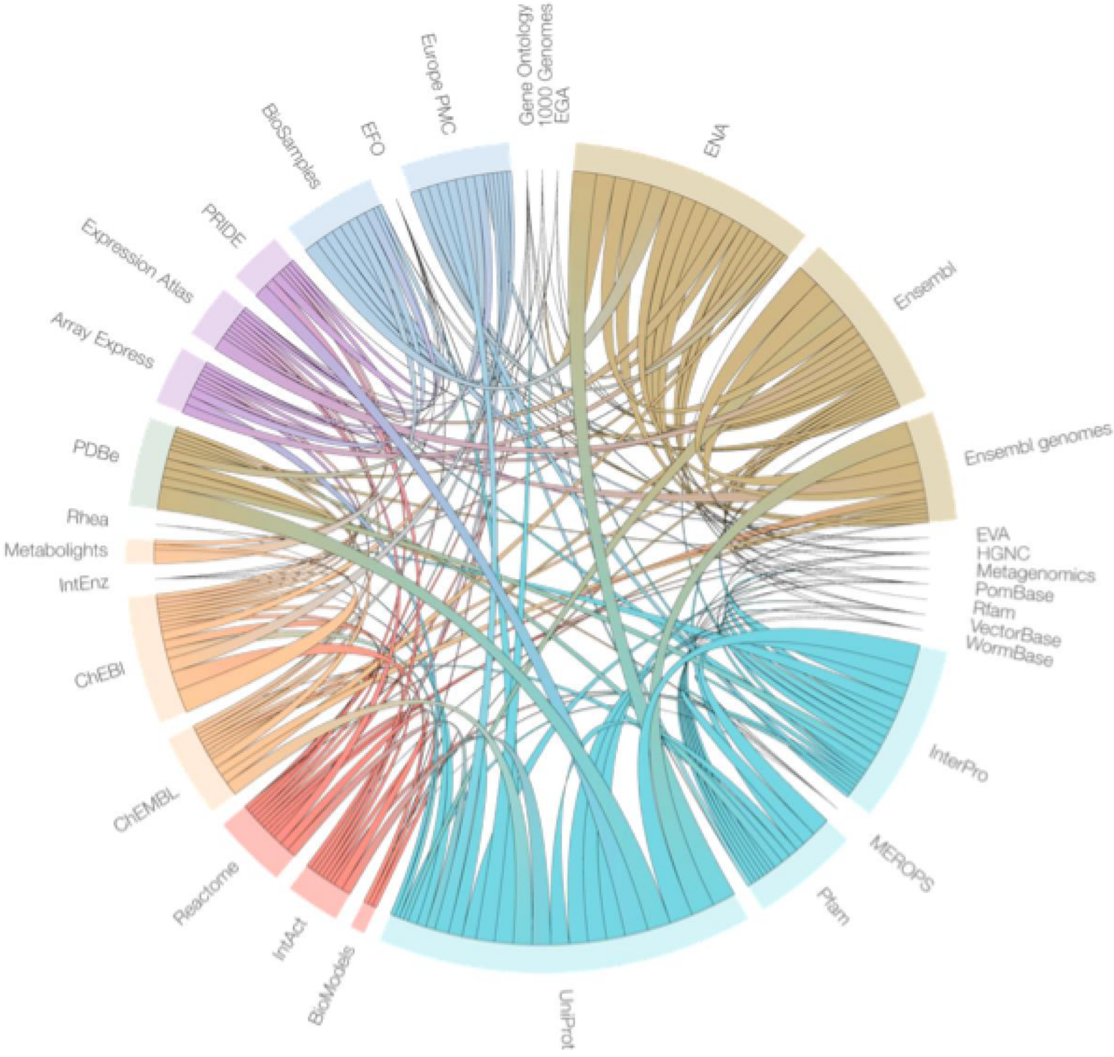

Figure 1: Cross-linking of major life sciences resources at EMBL-EBI (source: EMBL-EBI)

Aside from ORCID iDs, author records in INSPIRE include an INSPIRE ID and an author ID. The latter is generated automatically when a stub record is suspected of being a unique individual, while the former is created by cataloguers when they determine that the record belongs to a person. These IDs have little use outside of INSPIRE, so they are not pushed to ORCID as alternate identifiers.

The CERN Analysis Preservation[20] (CAP) service, currently under development, will include in its meta-data any known ORCID iDs. Authentication via ORCID could be possible with CAP, but as it is an internal tool, it is subject to authentication via CERN's login scheme. It is yet to be decided how best to use ORCID authentication in conjunction with necessary internal access restrictions. This decision depends on next steps elsewhere in CERN and on ORCID uptake across the organisation.

[20] https://analysis-preservation.cern.ch/ [not available outside CERN]

2.4.2 Identifier Linking and Versioning

In high energy physics (HEP), there is a provenance pipeline extending from the data originating in the experiments, passing through various stages of reconstruction and processing, and ending at figures included in publications. Generally speaking, the data volume is initially extremely large and narrows in scope as it nears publication. Since multiple researchers with multiple research interests can make use of the same large starting dataset, versions operate more as trees than in a strictly linear form: multiple datasets can be "versions" of another dataset, without relating to each other linearly. For this kind of data, the relationships most relevant from a DataCite Metadata Schema perspective would be those defined by the properties "IsDerivedBy" and "IsSourceOf".

Thus far, most of the data in this provenance pipeline have not been consistently associated with persistent identifiers, largely because the majority of it has not resided in repositories. An exception is HEPData[21], a database for the publication-ready tables of scattering data that are found at the narrow end of the HEP provenance pipeline. In HEPData, each record consists of the tables that accompany a single publication. Each table has its own DOI, as does the record for the package of tables. In turn, the record is associated with the original publication through the inclusion of the publication's DOI, arXiv ID and INSPIRE ID.

The main role of the long-standing INSPIRE literature database product is that of an aggregator. This means it needs to support the existence of multiple versions of publications, given the heavy reliance on preprints in HEP and the broader physics community. A master record exists for each paper as a concept; the various versions of a paper, whether preprints or final journal publications, are then linked from that master record. This grouping happens by DOI: an arXiv preprint gets linked to the final paper by virtue of arXiv including the DOI of the final published version in the preprint record. INSPIRE displays the most recent version of an arXiv paper to the user. Older versions are kept for archival purposes, but they are not shown to users (for example in searches). INSPIRE can accept data submissions, but this service is intended to serve the long tail of HEP data, i.e. that which is quite small and not affiliated with a publication. The low data submission volume permits manual minting of DOIs, allowing for responsive versioning on a case-by-case basis, either appending or minting a new DOI depending on user preference and severity of version change.

For other CERN scientific information products, the way data is versioned has not yet been standardised. For CERN Analysis Preservation (CAP), data versioning is still an open question, and potential solutions will be discussed as part of the THOR project. This new service is facing new challenges, as its content is very dynamic and connected to multiple other resources, and it should enable data citation while offering restricted access to resources while work is in progress.

The Zenodo-GitHub[22] integration enables researchers to take a snapshot of their software from the GitHub code repository and preserve it on Zenodo. Each code release (as implemented in the GitHub service) has its own new DOI assigned. The metadata sent to DataCite currently does not include a link to the repository as a **relatedIdentifier**, and Zenodo is not issuing a DOI for the collection of all releases/ versions.

[21] https://hepdata.net

[22] https://zenodo.org/account/settings/github

2.4.3 Organisational Identifiers

CERN is currently not making use of a standard organisational identifier scheme, but discussions are ongoing. Meanwhile, CERN is developing machine learning solutions to cluster and disambiguate reported organisational names within their products.

2.5 British Library

The British Library[23] is the national library of the United Kingdom, and is coordinating the EC-funded THOR project. It provides an extensive set of data and metadata services. Following regulators changes in 2013, it gathers digital content published in the United Kingdom[24] including the UK web domain. It also supports a national resource for theses, discussed below. The EThOS (e-Theses online service) holds records for 90% of all theses ever published in the UK, of which 40% are held in digital form.[25]

2.5.1 Versioning

EThOS, the e-theses online service at the British Library[26], aggregates records of UK doctoral theses. In some cases, separate copies of the thesis represented by one record could be held:

- By the BL (an electronic copy)
- And/or by the institution who awarded the thesis (in physical and electronic copy)
- And/or by the commercial organisation who digitised the thesis on behalf of the awarding institution (electronic copy)

In FRBR[27] (Functional Requirements for Bibliographic Records, a conceptual entity–relationship model for bibliographic data) the electronic copies and the hardcopies are separate manifestations of the work (the work being the intellectual content of the thesis). This model is illustrated in Figure 2. Discussions so far have focused on a single DOI for the electronic manifestation of the thesis and not for individual items (for example one for the British Library copy, one for the institutional electronic copy, one for the commercial copy). As such, this means discussions have looked at who should assign the DOI and when, and how the DOI can then be passed on in the metadata to aggregators.

Previous work in the Unlocking Thesis Data project (Grace, Whitton, Gould, & Kotarski, 2015) recommended the reservation of a DOI before the work had been completed. If an identifier is assigned before a hardcopy thesis is printed, it would also be applied to the hardcopy; it would then be a DOI for the work. The thesis as a work does not change content between electronic and hard copies, so it makes sense that only one identifier is required, and at the work level. However, there is one circumstance where a second identifier may be required for separate manifestations of the thesis: for theses

[23] http://www.bl.uk/

[24] http://www.legislation.gov.uk/uksi/2013/777/made

[25] http://ethos.bl.uk/

[26] http://ethos.bl.uk/

[27] https://www.loc.gov/cds/downloads/FRBR.PDF

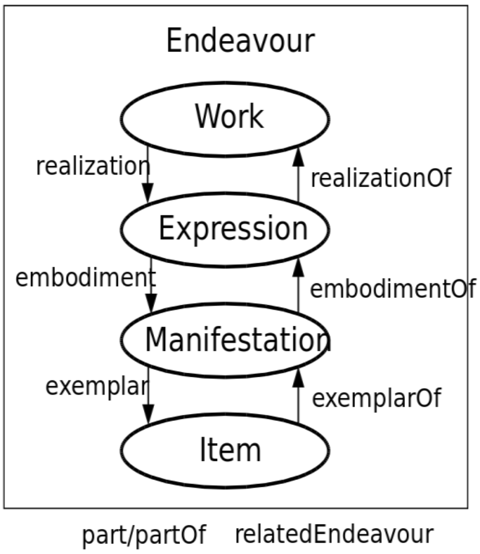

Figure 2: Basic Group 1 entities and relations of the FRBR model (RDF version) (Wikimedia Commons, 2007)

containing redactions. A thesis with redacted information could change the way in which its content is analysed and reused, and so ideally a separate identifier would be used for the redacted and full manifestations of a thesis. If an identifier is assigned prior to the completion of a thesis, it is likely the redacted and full version will have the same identifier. This is a complication that will need to be investigated further.

Consideration of FRBR also raises the issue of whether 'version of record' applies to data and other kinds of resources – and whether only a 'version of record' should have a DOI. There are valid reasons for 'non-record-versions' to have an ID, which raises the question of whether there needs to be a way to highlight which is a 'definitive' copy or version of a work.

In any instance where there are multiple versions of a resource, an issue arises for ORCID and other name identifiers: an author can select which record to claim when adding works to their records manually, and they are unlikely to claim multiple items of a work they have contributed to intentionally. There is potential with repositories and data centres being able to push items into author records that the same work will be present in a record multiple times, each with a different identifier, and authors will then need to manually remove them.

2.5.2 ISNI as Institutional Identifier

The British Library is one of the founding members of the ISNI International Agency[28]. The International Standard Name Identifier (ISNI) has leveraged the work of organisations that are responsible for uniquely identifying organisations in information resources, such as the library collections. If resources such as books or articles are written by or about an organisation, it is probable that the organisation already has an ISNI. ISNI already holds records for more than 500,000 organisations.

[28] http://isni.org

Since Ringgold was the first Registration Agency to offer ISNIs for organisations, its data has been loaded into the ISNI database as a primary source for organisation identifiers. This means that even without a match, a unique Ringgold ID will be assigned an ISNI. This has caused an additional level of duplication where a Ringgold ID has failed to find a match but in fact another ISNI has already been assigned for the same entity.

Any ISNI member or Registration Agency may edit or create ISNI identifiers for organisations of interest to them. Where a conflict is resolved by manually merging duplicates, the current rule would retain the ISNI associated with a Ringgold ID and deprecate the other one(s) merged. All deprecated ISNIs are retained in the ISNI record for that organisation so that former ISNIs will always resolve on the central database. Merges and splits that cause an ID to change will always be communicated to all ISNI member databases whose source code is associated with the ISNI record.

Issues with duplication in ISNI are reflected in way that institutional ID provider GRID[29] uses ISNIs in its database. These have been harvested from the ISNI public database, rather than being obtained from the Assignment Agency using its algorithms. This means that there is no reporting mechanism from the ISNI database to GRID as duplicates are merged and ISNI records improved.

It should be noted that ISNI hold organisational IDs from many sources, not just Ringgold. This means that a different ISNI registration agency may have already created an ISNI record for a given organisation, leading to duplicates within ISNI that are not connected. This is in part due to the ISNI matching algorithm being publication based. As an example, the ISNI ID for the European Commission is sourced from 11 authorities, including VIAF and the Library of Congress name authority file/NACO, but Ringgold is not one of them.

2.6 PANGAEA

PANGAEA, the Publishing Network for Geoscientific & Environmental Data[30], is an Open Access library that archives, publishes and distributes geo-referenced data about climate variability, the marine environment and geological research.

2.6.1 Identifier Linking

Each dataset can be identified, shared, published and cited by using a DOI, minted by DataCite. PANGAEA supports three versions of citable datasets:

- Data supplement, i.e. data that is supplementary to a scientific paper and is thus an integral part of the paper and of its peer-review
- Independent data publication not linked to the publication of an article
- Peer-reviewed data publications

[29] http://grid.ac

[30] http://www.pangaea.de/about/

The dataset metadata record also relates to an abstract, which is a brief natural language description of the dataset publication. For samples and measurements, the provision of position(s) is mandatory. The metadata record thus supports a relation to spatial coverage. The record also allows for relations to the project within as well as relations to the events during which the data was collected. The record holds an additional relation to detailed metadata about the relevant parameters. Finally, the dataset metadata record allows for a relation to citations of other research objects, which may be literature or (PANGAEA) datasets related to the published dataset.

PANGAEA attempts to resolve the ORCID iD of authors named in the metadata record. If the resolution succeeds, then author names are annotated with their ORCID iD. If the ID cannot be resolved then the email is shown instead (if known). The resolution is based on a heuristic whereby the algorithm associates an ORCID iD to an author if the author names (between PANGAEA and ORCID) match and one or more resource objects claimed by the author at ORCID are known to PANGAEA as being authored by the person.

Datasets link to a Uniform Resource Locator (URL), specifically a HTTP URI, which points to the location of the raw data file. The URL is a machine actionable linking identifier that enables the download of data related to a dataset.

Datasets deposited in PANGAEA will be linked automatically to corresponding articles in the Elsevier ScienceDirect service and can be linked to a Google Map displaying the geographical locations of PANGAEA datasets.

As data citations from literature are rare, PANGAEA is keeping track of the link from datasets back to articles – the “reverse links”. Another important linked resource maintained by PANGAEA dataset records is the CC license under which the dataset was published.

2.6.2 Versioning

Datasets published at PANGAEA cannot be modified after final DOI minting and publication. Different versions of a dataset published at PANGAEA are thus treated as new publications and individually obtain a DOI. DOIs of older and newer dataset versions are linked. The older version links to the DOI of the newer version via the relation type “New Version”, which corresponds to the “IsPreviousVersionOf” DataCite relation type. If the older version was deleted, its DOI links via the relation type “Replaced By” instead.

2.6.3 Relation Types

In dialogue with services such as DataCite, PANGAEA defines and utilises a number of relation types and mappings between its own types and the relation types of other schemas. Table 1provides an overview of PANGAEA relations types and their mapping to DataCite relations types. The semantics of the mapped terms are considered to be approximately equal.

Table 1: Mapping common PANGAEA and DataCite relation types

PANGAEA	DataCite
Related to	References
Supplement to	IsSupplementTo
New version	IsPreviousVersionOf
Replaced by	IsPreviousVersionOf
Child having "In: Parent dataset citation"	IsPartOf
Further details	IsDocumentedBy
Source data set	IsDerivedFrom
Other version	IsVariantFormOf

3 Conceptual Model for Persistent Identifier Linking

One major challenge in persistent identifier linking is scalability. The current implementations are too focused on linking specific identifier types. As a result, they do not scale well to a variety of different persistent identifiers or to different kinds of resources being linked together. A conceptual model for persistent identifier linking can help implement solutions that scale well and thus help with adoption of persistent identifier linking. This is particularly important in areas where persistent identifier linking is still at a fairly early stage, for example linking research outputs to institutions and funding information.

3.1 Linkage as Triples

In its simplest form, one persistent identifier is linked to another persistent identifier. The two persistent identifiers are the minimal required information. In addition, a way to store the linkage is also needed. In the Resource Description Framework (RDF) data model[31] this would be described as a triple in the form *subject-predicate-object*.

3.2 Describing the Relation

In practice, the relation between the two persistent identifiers is often implicitly characterised by the kinds of resources identified by them. For example, linking a grant identifier to a dataset identifier might implicitly indicate that the grant provided funding for creation of the dataset. Similarly, linking an ORCID iD to a dataset identifier might implicitly indicate that this contributor was involved in the creation of

[31] https://www.w3.org/TR/PR-rdf-syntax/

the dataset. Of course, many possible relationships between these types of objects may also exist. In the first example, the dataset could have been a key input to a project, rather than an output.

In some cases, the type of relation between two identifiers is further described in the metadata, in particular when two research outputs are linked together, for example **A cites B**, or **A IsNewVersionOf B**. The DataCite Metadata Schema, for example, provides a **relationType** controlled vocabulary to describe the relation type of two linked research outputs, and the **contributorType** controlled vocabulary to describe the relation type between contributor and research output. Other relation type standards are emerging, for example the CREDIT contributor roles taxonomy[32].

When persistent identifier **B** is included in the metadata of persistent identifier **A**, this implies that whoever submitted the metadata for persistent identifier **A** is describing the linkage. An example would be ORCID identifier http://orcid.org/0000-0002-4133-2218 in the metadata of dataset http://doi.org/10.1594/PANGAEA.733793, maintained by PANGAEA. All metadata in the DataCite Metadata Store is maintained by the respective data centre that registered the DOI. ORCID is using a different approach: entries in the ORCID registry can be supplied by multiple organisations, and the provenance information is made available via the **source** attribute.

The linkage between two persistent identifiers usually includes information about when the linkage was documented. This is either implicit (the publication date of the metadata record includes the linked identifier), or explicit (the date the linkage was documented is included). DataCite uses the former approach whereas ORCID uses the latter.

3.3 Persistent Identifiers as HTTP URIs

When persistent identifiers are referred to as HTTP URIs, as recommended by the Den Haag Persistent Object Identifier – Linked Open Data Manifesto (Knowledge Exchange, 2011), persistent identifier linking can become compatible with the RDF data model [1,33]. The persistent identifier expressed as an HTTP URI is globally unique and machine actionable, as recommended in the Joint Declaration of Data Citation Principles (Martone, 2014). In their display guidelines, ORCID[34], DataCite and CrossRef[35] recommend to express their persistent identifiers as HTTP URIs, for example http://orcid.org/0000-0002-1825-0097 or http://doi.org/10.5438/0010.

Expressing all identifiers as URI provides global uniqueness and makes them machine actionable, but poses the following challenges:

- Identifiers can be expressed as more than one URI
- The HTTP URI for the identifier might not be stable even if the identifier itself does not change
- Grouping of identifiers by identifier type becomes more difficult

[32] http://casrai.org/CRediT

[33] http://wifo5-03.informatik.uni-mannheim.de/bizer/pub/LinkedDataTutorial

[34] https://about.orcid.org/trademark-and-id-display-guidelines

[35] http://www.crossref.org/02publishers/doi_display_guidelines.html

The above issues can be addressed by using a resolving service, such as the Handle System used for DOI names; identifiers.org, used for life sciences accession numbers, is also recommended for linking persistent identifiers. This provides a more stable URI, independent of location changes of the resource. As a result, it can support multiple locations and multiple representations of a resource, ideally supporting content negotiation. This approach overcomes one major limitation of traditional HTTP URIs, namely that they may become permanently unavailable – so-called link rot, a particular problem for scholarly content (Klein et al., 2014). The ORCID service also supports content negotiation for ORCID identifiers expressed as HTTP URIs. The quality control of a resolving service should include regular checks to ensure that all persistent identifiers are correctly resolving to a URL location.

Many life sciences databases use accession numbers, identifiers that are only unique within the context of the specific database. While robust linking between these life sciences databases, taking the database context into consideration, has been in place for many years, efforts are also underway to establish a more generic approach to identifier linking, using the resolving service identifiers.org provided by EMBL-EBI.

Multiple identifier schemes are common for some resources: PMID and DOI are often combined for journal articles, accession number and DOI for some life sciences resources, and ORCID iD and ISNI for contributors. Providing linkage between multiple identifiers for the same resource, for example as **alternateIdentifier** in the DataCite metadata, facilitates cross-linking across identifier schemes.

More work is needed to express unique identifiers for some scholarly resources as HTTP URIs, including organisations or funding information. Part of the challenge is in handling unique identifiers that can only be used in a closed system (for example a proprietary service that requires a subscription). Ideally all linking identifiers should resolve to a publically accessible landing page.

3.4 Centralised Infrastructure for Identifier Linking

ORCID and DataCite collect information about identifier linking in central locations, the ORCID registry and DataCite Metadata Store, respectively. Linking information is provided using standardised metadata. This greatly facilitates discovery of linked identifiers and complements the distributed nature of Linked Open Data[3].

3.5 Further Work

Expressing persistent identifiers as URIs is adequate when all identifier types are known in advance and resolution mechanisms are considered stable. Yet it becomes an issue when identifiers can be resolved in more than one manner, as, for example, is common for EMBL-EBI resources.

There are identifiers in the wild that, while differing in resolvable representation, identify the same conceptual entity. A notable example of this can be seen with PDB Identifiers mentioned earlier in this document. For a given identifier of this type, there are multiple ways of resolving it. From a certain perspective, then, non-equivalent URIs are in fact equivalent persistent identifiers. For example, the identifier **3coj** may be resolved as follows:

- PDB Europe: http://www.ebi.ac.uk/pdbe/entry/pdb/3coj
- PDB Japan: http://pdbj.org/mine/summary/3coj
- RCSB Protein Data Bank: http://www.rcsb.org/pdb/explore/explore.do?structureId=3coj
- Protopedia: http://proteopedia.org/wiki/index.php/3coj
- PDBsum: https://www.ebi.ac.uk/thornton-srv/databases/cgi-bin/pdbsum/GetPage.pl?pdbcode=3coj
  http://www.rcsb.org/pdb/explore/explore.do?structureId=3coj

The resolving service identifiers.org knows of all five of these locations, and redirects to one of them based on service uptime, user preference, and other criteria.

For systems that group entities by identifier such as ORCID, this presents a problem. In normal operation, all versions of metadata for an object are grouped around an external identifier, such as a DOI. This means that multiple versions asserted by different third-parties can be treated as a single entry in an ORCID record. For example, Figshare may add a work to an ORCID record, DataCite auto-update could add the same work, as could an institutional CRIS system or the researchers themselves.

When dealing with identifiers like PDB, it would be possible to use specialised processing to match equivalent identifiers, but this presents scalability problems and may be useful only when using established resolving services such as DOI, Handle or possibly the newer identifiers.org. This requires the ability to query these resolving services in the reverse direction, such that, for example, a query for http://www.ebi.ac.uk/pdbe/entry/pdb/3coj points to http://identifiers.org/pdb/3coj.

Another route is to treat PDB Europe identifiers as being conceptually as different from PDB Japan as DOIs are from Handles. This benefits from simplicity but effectively ignores the problem, providing no way of associating the two identifiers.

Explicit support of global uniqueness and machine actionability via HTTP URIs may obviate the need for controlled lists of allowed identifier types (as, for example, used by ORCID and DataCite) in most cases. Controlled vocabularies can, in fact, impede interoperability by restricting links to specific systems. To address this, ORCID will be moving to a system whereby the vocabulary is understood and defined, yet still extensible. This will enable new identifier types and their associated metadata (for example multi-language descriptions) to be added to the registry in response to community needs. Thought will be given on how these identifier types relate to one another and the representation of equivalence between similar identifiers. The applicability of the RDF concepts of ontologies and classes will be evaluated as part of this work.

In addition, work will be done so that identifiers that are not expressed as URIs can be expressed as such, and vice versa, by requiring a method of transformation in the identifier metadata. Another ORCID goal for 2016 is to expose ISNI identifiers alongside Ringgold in the ORCID API to aid interoperability.

For the next revision of the DataCite Metadata Schema, DataCite will evaluate whether identifier linking in the metadata can be further harmonised, using the same concepts when linking to data, contributors, funding information, institutions, and so on. This includes a re-evaluation of the use of controlled lists, for example for the **relatedIdentifier** field.

DataCite is currently collecting information about linked identifiers solely from its data centres. Work is ongoing to complement these links collected in the DataCite Metadata Store with links provided by external services (such as data citations provided via CrossRef) in a new DataCite Link Store[36]. This will be enacted using a model that resembles how the ORCID Registry operates.

4 Discussion

In the previous sections we have described how ORCID, DataCite and data centres from four different disciplines have implemented persistent identifier linking. We have introduced a conceptual model of persistent identifier linking as a way to make persistent identifier linking more generic and thus scale better. In the discussion, we now look at how this work can inform the following important use cases for persistent identifier linking:

- Linking identifiers for versioning data
- Linking identifiers for articles and data
- Linking data with contributors, institutions and funders

4.1 Identifier Linking for Versioning Data

The ability to refer to a specific version of a dataset is an important use case, and is one of the recommendations of the Joint Declaration of Data Citation Principles (Martone, 2014). Version information, if available, should be part of a data citation. At the same time, there is currently not yet a community consensus on how to do versioning of data; the decision on how to do data versioning is ultimately made by the data repository:

- Versioning of data is important for specificity and verifiability
- The data repository is ultimately responsible for decisions about versioning

Given that versioning is dependent on the context and respective community practices, a single set of specific recommendations for data versioning cannot be made. What is possible, however, is to recommend a set of best practices, based on current implementations for data versioning:

1. **Major version** changes require a new persistent identifier and new set of metadata, whereas for **minor version** changes only the data and/or metadata are updated; the persistent identifier does not change.
2. A naming convention for the persistent identifier should not be the only place where version information is encoded.
3. Both the version number and related identifiers of other versions can be described in the metadata.
4. Both the version number and related identifiers of other versions can be included in the landing page.

[36] https://ls.datacite.org

5. Humans and machines should be able to easily see multiple versions if they exist, and be able to tell whether they are looking at the newest version of a dataset.
6. Data and metadata of older versions should be kept available if possible, using a tombstone page if the data are no longer available.
7. Information about what changed in comparison to the previous version is desirable.
8. A collection that includes all versions of a dataset can be assigned a persistent identifier and aggregate their version information.

4.1.1 Major vs. Minor Versions

Although data centres may have different criteria about whether to label a version change as minor or major, there is general agreement that the distinction between minor and major version changes is useful. This distinction is often implemented by requiring a new persistent identifier for major versions. A common practice is to describe minor versions with an appended version number (for example version 1.0 is updated to version 1.1). Major versions are described with an incremented version number (for example version 1.0 is updated to version 2.0).

In general, basic metadata changes that do not affect the citation are considered minor. Changes to the data, including addition of files, are considered major (Data Science at The Institute for Quantitative Social Science, nd; UK Data Archive, 2014).

4.1.2 Landing Pages

Landing pages are important for proper versioning of data. Starr et al. (2015) recommend that persistent identifiers for data resolve to a landing page that contains metadata and other relevant information about the dataset, and links to the dataset itself. The landing page should include version information and links to other versions of the same dataset. Van de Sompel et al. (2014; 2015) highlight the need for landing pages to be machine readable rather than only focused on human users, which can complement the machine readable access to metadata provided by DOI content negotiation[37].

In addition to a landing page for each individual version of a dataset, a landing page that summarises all versions of a dataset is recommended. This landing page should be associated with a persistent identifier for the collection of dataset versions.

Tombstone pages should be maintained for versions of datasets that have been removed, although removal is generally discouraged. Dataverse, for example, deaccessions data only when “legally compelled” (Dataverse Project, nd). In some cases the provider of the persistent identifier might assist with maintaining tombstone pages, for example when the data centre ceases to exist, or when versions of a dataset are stored in multiple data repositories.

[37] http://crosscite.org/cn/

4.1.3 Future Work: Notifications

There is currently no standard mechanism in place to notify users that a new version of a dataset has been made available. CrossRef is providing such a service for their DOI names with the optional CrossMark[38] service. There is clearly a use case for notifications, in particular for users citing or otherwise reusing a dataset. These notifications could go to the user, but also to the publisher or data centre who published the article or dataset referencing a dataset with a new version. The THOR project will explore notifications of new versions as part of its future work, as CERN and other THOR partners are interested in this functionality.

4.1.4 Future Work: Dynamic Datasets

Dynamic datasets present specific challenges that will be addressed in future THOR work on accessing data using new resolution methods (starting in June 2016).

4.2 Identifier Linking for Data and Articles

DataCite metadata captures links from datasets to persistent identifiers for articles via its **relatedIdentifier** field. Data centres that do not use DOI names as persistent identifiers use similar approaches. Machine actionable persistent identifier linking is an important part of the Joint Declaration of Data Citation Principles (Martone, 2014). The most commonly used persistent identifier for scholarly articles is the DOI. Using DOIs to describe citations between articles is a well-established practice and the core mission of the CrossRef DOI registration agency.

Data citation by articles is a relatively new concept, and there is not yet a standard practice of how these data citations are described in metadata. One consequence is that many journal articles do not cite data, and data identifiers are not part of the article repositories' metadata. They can only be found with access to the full text article. This is, for example, a common pattern with life sciences accession numbers. Extracting data citations from full text via text mining is labour-intensive and brittle, and – in contrast to metadata – fails without access to and permissions for reuse of the full text article.

The conceptual model of identifier linking described in this report can help establish better practices for data citation. The model stipulates that we:

- Use persistent identifiers for both the article and the dataset in a data citation
- Express these persistent identifiers as HTTP URIs (rather than, for example, using accession numbers)
- Optionally, describe the relationship between the article and the data; for example, to discriminate data that the article results are based on, from data that are cited because they are related, but have not directly been fed into the results
- Use centralised infrastructure to make it easier to find these article/data links; for example, the CrossRef and DataCite metadata search

[38] http://www.crossref.org/crossmark/

The Research Data Alliance/World Data Service (RDA/WDS) Data Publishing Services Working Group[39] is working on implementing a service infrastructure for data citations. In a workshop in January 2016, it decided to build a service infrastructure that is consistent with the persistent identifier linking conceptual model, for example using triples and expressing persistent identifiers as URI. THOR partners PANGAEA, ANDS, Elsevier, EMBL-EBI and DataCite are participating in this work, as are other organisations, including OpenAire and CrossRef. The THOR project will start work on data citation services based on this conceptual model in 2016.

4.3 Data Linking with Contributors, Institutions and Funders

4.3.1 Contributors

Linking data and contributors is the focus of the THOR project. Persistent identifier linking of ORCID iDs for contributors with DOI names for data is implemented in both the DataCite Metadata Schema and ORCID Schema. The first persistent identifier linking service between ORCID and DataCite was implemented in 2013 as part of work in the ODIN project. In the pilot service, contributors were able search DataCite metadata[40] via a search interface and claim the DOI and metadata to their ORCID record. As part of the work in the THOR project, DataCite will implement this “search and link” functionality in the default DataCite search. A pre-release implementation is available[41] as of August 2015.

The “search and link” functionality depends on self-claims by researchers after a dataset has been published. A different approach to identifier linking focuses on adding ORCID identifiers to the DOI metadata when the dataset is created. DataCite can then automatically forward the ORCID iD/DOI link to the ORCID Registry. ORCID and DataCite are working on implementing this so-called “auto-update” workflow[42].THOR partners EMBL-EBI, PANGAEA and CERN are improving their workflows on adding ORCID iDs on dataset submission in this way.

Adding datasets to the ORCID Registry that do not use DOI names as persistent identifiers is currently limited by the fact that ORCID uses a controlled list of allowed identifier types that are encoded within the XML schemas that define the ORCID service interface. This is obviously not a scalable solution as every new identifier type that is added changes the schema and therefore introduces breaking changes. It also pushes the burden of configuration to developers, when it should be in the hands of those that manage the service. This makes responding to requests for new identifier types slower than ideal. What is required is a mechanism to add references to identifier types in a controlled yet flexible way.

One solution is to modify the ORCID registry to enable client applications to dynamically add identifier types to the registry and to remove type enumerations from the schema definitions. This would remove the reliance on XML schemas for validation and passes the responsibility onto the service itself. A

[39] https://rd-alliance.org/groups/rdawds-publishing-data-services-wg.html

[40] http://datacite.labs.orcid-eu.org

[41] https://search.labs.datacite.org

[42] https://orcid.org/blog/2015/10/26/auto-update-has-arrived-orcid-records-move-next-level

separate part of the API would then be responsible for the querying, creation and updating of identifier types. Furthermore, features such as identifier validation and URL mapping could be provided as part of this API.

Another solution is to resolve identifier types through a third party, much like identifiers.org currently does. All "non-standard" identifier types could be routed through this service. However, it must be noted that this adds a dependency on a third party that must be considered sustainable to prevent link-rot of the identifier resolution.

4.3.2 Institutions

Although institutional information is available in many DataCite DOI metadata records via the **HostingInstitution** field, only in some cases are institutional identifiers provided. One reason is that uptake of organisational identifiers lags behind uptake of identifiers for contributors and data, as described in our previous report (Fenner et al., 2015).

In that report, we identified an additional requirement: organisational identifiers should be expressed as HTTP URIs that resolve to a publicly available landing page with human and machine readable information. ORCID and DataCite are co-organising a workshop on organisational identifiers at the Force16 conference[43] in April 2016 to advance these issues.

As the ORCID Registry has been using Ringgold and ISNI organisational identifiers for some time, it is possible to do a persistent identifier "cross-walk" via the ORCID iD as intermediary to link persistent identifiers for data with persistent identifiers for institutions.

In the past few months, we have seen the emergence of a new provider of organisational identifiers in the form of GRID, the Global Research Identifier Database[44]. The GRID database contains identifiers from both ISNI and the CrossRef Funder identifier database, as well as other sources. This would provide another, and more direct, solution to linking organisation identifier types through a third party.

4.3.3 Funding Information

Open Funder Registry IDs have been available for some time, but grant numbers were not included in the funding information that could be encoded in DataCite metadata. This will change with the release of version 4.0 of the DataCite Metadata Schema later in 2016. This change will require DataCite data centres to implement and adapt the collection of funding information in the data submission workflow.

Similar to the "search and link" and "auto-update" workflow options for linking contributors and data, we could also implement a "search and link" workflow for funding information. This would complement the "auto-update" option of adding funding information during the data submission workflow with the "search and link" option of claiming funding information retroactively. This would allow researchers to add funding information to their works after they have been published, serving as a temporary solution until "auto-update" workflows for funding information are fully in place at data centres.

[43] https://www.force11.org/meetings/force2016

[44] http://grid.ac/

As discussed in our previous report (Fenner et al., 2015), ORCID links funding information to contributors whereas DataCite (and CrossRef) link funding information to research outputs, representing different views on funding. Further work is needed to reconcile these different persistent identifier linkages.

5 Conclusions

This report describes an important function of persistent identifiers: persistent identifier linking. It looks at how this is currently implemented at ORCID, DataCite and repositories in four different disciplines. While the importance of persistent identifier linking for scholarly e-Infrastructure is clearly understood, the adoption in several areas, such as linking identifiers for data with funding information, is still at an early stage.

We identified many small hurdles that need to be overcome, and developed a conceptual model for persistent identifier linking to overcome them. Our model supports implementation of scalable solutions and addresses important use cases for scholarly e-Infrastructure, including data citation.

6 References

Cook, C. E., Bergman, M. T., Finn, R. D., Cochrane, G., Birney, E., & Apweiler, R. (2015, December 15). The European Bioinformatics Institute in 2016: Data growth and integration. Nucleic Acids Res. Oxford University Press (OUP). http://doi.org/10.1093/nar/gkv1352

DataCite Metadata Working Group. (2014). DataCite Metadata Schema for the Publication and Citation of Research Data v3.1. DataCite e.V. Retrieved from http://doi.org/10.5438/0010

Data Science at The Institute for Quantitative Social Science. (nd). Dataset + File Management — Dataverse.org. (2016, January 19). Dataset + File Management - Dataverse.org. Retrieved February 29, 2016, from http://guides.dataverse.org/en/latest/user/dataset-management.html

Dataverse Project (nd). Harvard Dataverse Preservation Policy | The Dataverse Project - Dataverse.org. Retrieved February 3, 2016, from http://dataverse.org/best-practices/harvard-dataverse-preservation-policy

EMBL-EBI. (nd). Services A to Z. International Nucleotide Sequence Database Collaboration. Retrieved from http://www.ebi.ac.uk/services/all

Fenner, M., Demeranville, T., Kotarski, R., Vision, T., Rueda, L., Dasler, R., … THOR Consortium. (September 2015). D2.1: Artefact, Contributor, and Organisation Relationship Data Schema. Zenodo. Retrieved from http://doi.org/10.5281/ZENODO.30799

Grace, S., and Whitton, M., and Gould, S., and Kotarski, R. (2015). Mapping the UK thesis landscape: Phase 1 project report for Unlocking Thesis Data. Project Report. University of East London, London. (10.15123/PUB.4307).

Klein, M., Van de Sompel, H., Sanderson, R., Shankar, H., Balakireva, L., Zhou, K., & Tobin, R. (2014). Scholarly context not found: one in five articles suffers from reference rot. PLoS ONE, 9(12), e115253. http://doi.org/10.1371/journal.pone.0115253

Knowledge Exchange. (2011). Den Haag Persistent Object Identifier – Linked Open Data Manifesto. Retrieved from http://ke-archive.stage.aerian.com/admin/public/dwsdownload.aspx%3Ffile=%252ffiles%252ffiler%252fdownloads%252fpersid%252fworkshop+poid%252fden+haag+manifesto+20110825_2.pdf

Martone M. (ed.). (2014). Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. San Diego CA: FORCE11; 2014. Retrieved February 29, 2016, from https://www.force11.org/group/joint-declaration-data-citation-principles-final

ODIN Consortium, Fenner, M., Thorisson, G., Ruiz, S., & Brase, J. (2013). D4.1 Conceptual model of interoperability. Figshare. Retrieved from http://doi.org/10.6084/M9.FIGSHARE.824314

Starr, J., Castro, E., Crosas, M., Dumontier, M., Downs, R. R., Duerr, R., … Clark, T. (2015). Achieving human and machine accessibility of cited data in scholarly publications. PeerJ Computer Science, 1, e1. Retrieved from http://doi.org/10.7717/peerj-cs.1

UK Data Archive. (22 May 2014). Preservation Policy. Version: 08.00. University of Essex. Retrieved from http://www.data-archive.ac.uk/media/54776/ukda062-dps-preservationpolicy.pdf

Van de Sompel, H., Sanderson, R., Shankar, H., & Klein, M. (2014). Persistent Identifiers for Scholarly Assets and the Web: The Need for an Unambiguous Mapping. International Journal of Digital Curation, 9(1), 331–342. Retrieved from http://doi.org/10.2218/ijdc.v9i1.320

Van de Sompel, H., & Nelson, M. L. (2015). Reminiscing About 15 Years of Interoperability Efforts. D-Lib Magazine, 21(11/12). Retrieved from http://doi.org/10.1045/november2015-vandesompel

Wikimedia Commons (2007). File:FRBR-Group-1-entities-and-basic-relations.svg - Wikimedia Commons. Retrieved from https://commons.wikimedia.org/wiki/File:FRBR-Group-1-entities-and-basic-relations.svg

Appendix A: Project summary

The **THOR** project establishes a sustainable international e-infrastructure for persistent identifiers that enables long-term access to critical information about the life cycle of research projects. It enables seamless integration between articles, data, and researcher information creating a wealth of open resources. This will result in reduced duplication, economies of scale, richer research services, and opportunities for innovation.

The project has four concrete aims:

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

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

The consortium will develop systems to embed new PID resolution techniques into existing services to support seamless direct access to artefacts, and in particular data. We will create services to allow associations between datasets, articles, contributors and organisations at the time of submission. Building on these, we will deliver the means to integrate trans-disciplinary PID services in community-specific platforms, focussing on cross-linking, claiming mechanisms and data citation (guided by the FORCE 11 data citation principles[45]).

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

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

Appendix B: Terminology

Additional terms are defined below:

Term	Definition
ANDS	Australian National Data Service
API	Application Programme Interface
arXiv	Open access e-print archive (Physics, Mathematics, Computer Science, Quantitative Biology, Quantitative Finance and Statistics)
BL	British Library
CERN	European Organisation for Nuclear Research
ChEMBL	European Bioinformatics Institute
CRIS	Current Research Information Systems
CrossRef	Digital Object Identifier Registration Agency working to make content easy to find, link, cite and assess in scholarly publishing.
DataCite	An organisation that develops and supports methods to locate, identify and cite data and other research objects. Specifically, DataCite develops and supports the standards behind persistent identifiers for data, and the members assign them. See https://www.datacite.org
Dataverse	Open source research data repository framework
DOI	Digital Object Identifier
DRYAD	Dryad Digital Repository: curated resource that makes the data underlying scientific publications discoverable, freely reusable, and citable.
ELSEVIER	Academic publishing company that publishes medical and scientific literature
EMBL-EBI	European Bioinformatics Institute , part of the European Molecular Biology Laboratory
HEP	High Energy Physics
ID	Identifier
INSDC	International Nucleotide Sequence Database Collaboration
ISNI	International Standard Name Identifier
ODIN	ORCID and DataCite Interoperability Network
ORCID	An organisation that creates and maintains a registry of unique researcher identifiers and a transparent method of linking research activities and outputs to these identifiers. See http://orcid.org
PANGAEA	Publishing Network for Geoscientific & Environmental Data
PDB	Protein Data Bank
PID	Persistent Identifier
PLOS	Public Library Of Science
PMID	Unique identifier number used in PubMed
RDF	Resource Description Framework
URI	Uniform Resource Identifier
VIAF	Virtual International Authority File