# ODIN: the ORCID and DataCite interoperability network

Martin Fenner

Public Library of Science,
San Francisco, CA, USA
Email: mf@martinfenner.org

Laurel L. Haak

ORCID, Bethesda, MD, USA
Email: l.haak@orcid.org

Gudmundur A. Thorisson

deCODE Genetics, Reykjavik, Iceland
Email: gthorisson@gmail.com

Sergio Ruiz*

ECMWF, Reading, UK
Email: sergioruizperez@gmail.com
*Corresponding author

Todd J. Vision

Department of Biology,
University of North Carolina at Chapel Hill, USA
Email: tjv@bio.unc.edu

Jan Brase

German National Library of Science and Technology,
Hannover, Germany
Email: jan.brase@tib.uni-hannover.de

**Abstract:** Research data is increasingly seen as the most significant untapped resource in scholarship. Awareness and practice of referencing and citing research data is increasing, and different initiatives to unambiguously identify datasets are in place. Steps are being taken to identify the individuals who created or contributed to research outputs. Lack of interoperability between the different initiatives to identify datasets and contributors remains a major hurdle. The ODIN project (*ORCID and DataCite Interoperability Network)* tries to address this need. ODIN builds on the ORCID and DataCite initiatives to uniquely identify scientists and data sets and connect this information across multiple services and infrastructures. It aims to address some of the critical open questions in the area. We describe a conceptual model to solve the interoperability between different identifiers for data and people.

**Keywords:** research data; authors; ODIN; ORCID; DataCite; interoperability; DOI; persistent digital identifiers; datasets; scholarly communication.

**Reference** to this paper should be made as follows: Fenner, M., Haak, L.L., Thorisson, G.A., Ruiz, S., Vision, T.J. and Brase, J. (2014) 'ODIN: the ORCID and DataCite interoperability network', *Int. J. Knowledge and Learning*, Vol. 9, No. 4, pp.305–325.

**Biographical notes:** Martin Fenner is the technical lead for the PLOS Article-Level Metrics project. Before taking this position in 2012 he worked as a medical oncologist at the Hannover Medical School Cancer Center in Germany. He has served on the ORCID Board from 2010 to 2012 and is a member of the ORCID Outreach Steering Group.

Laurel L. Haak, ORCID Executive Director. She drives awareness of the ORCID mission, building strategic relationships, working with a broad range of constituents, ensuring organisational persistence, and directing ORCID staff and contractors. Previously, she was Chief Science Officer at Discovery Logic, Inc.; a program officer for the US National Academies' Committee on Science, Engineering, and Public Policy; and editor of Science's Next Wave Postdoc Network at the American Association for the Advancement of Science. She received a BS and an MS in Biology from Stanford University and a PhD in Neuroscience in 1997 from Stanford University Medical School, and she was a postdoc at the US National Institutes of Health.

Gudmundur A. Thorisson, Academic and Tech Head interested in Bioinformatics, open access, Linked Data publishing and scholarly communication. He has been involved in various projects relating to identity & unique identifiers in research, including ORCID and VIVO. He holds a PhD from the University of Leicester. Before starting postgraduate studies in the United Kingdom in 2006, he spent a number of years working as a scientific programmer in industry and academia, after graduating with a BSc in Biology from the University of Iceland. He currently works on making sense of large-scale genomics data for research firm deCODE Genetics in Reykjavik, Iceland.

Sergio Ruiz is an Information Scientist. At present he is working for ECMWF (www.ecmwf.int) as Communications and Business Manager. From 2013 to 2014 he worked for DataCite as Operations Officer. Among other tasks he took care of the coordination of the ODIN Project. From 2009 to 2012 he worked for CERN (Geneva, Switzerland), where he took care of the coordination of two European Projects (SOAP and ODE); working on their administrative, financial, legal and scientific aspects. His Library and Information Science degree is complemented by a post-graduate degree in Scientific Information and Communication and a Master in Business Administration.

Todd J. Vision is an Associate Director for Informatics at the National Evolutionary Synthesis Center, and Associate Professor of Biology at the University of North Carolina at Chapel Hill, and a co-founder of the Dryad Digital Repository.

Jan Brase graduated in Mathematics at the university of Hannover in 1999 doctor in Computer Science in 2005. Since September 2006 he coordinated the DOI registration agency at TIB and since January 2010 to December 2014 he was Managing Director of DataCite. He was chair of the International DOI Foundation (IDF), president of the International Council for Scientific and Technical Information (ICSTI) and Co-Chair of the CODATA Task group on Data Citation. In 2011 he received the German 'Library Hi-Tech award'. At present he is EU Research Liaison Officer at German National Library of Science and Technology (TIB).

## 1 Introduction

The goal of the ODIN project is to identify and resolve issues relating to the 'missing thin layer' of persistent identifiers needed for a globally connected and interoperable scholarly communication e-infrastructures. This project aims to support and stimulate the adoption of interoperable identifiers for researchers, research works and their outputs, namely publications and data. In addition, it facilitates the information flow between research communities, leading to greater re-use of data and innovative exploitation of the existing knowledge.

ODIN focuses on the integration and scalability of the DataCite and ORCID persistent identifier initiatives, and ultimately will provide a roadmap for tackling four main challenges concerning research data: *accessibility*, *discovery*, *interoperability*, and *sustainability*.

### *1.1 Identifying, discovering and citing data*

Data are a major untapped resource in scientific research. Research data have the capacity to engender insights that will lead to entirely new products, services, and solutions to the world's grand challenge problems. Unfortunately, today we lack the infrastructure to realise the benefits of that capacity. Useful free flow of data currently is currently not possible. This is not because of the absence of computer networks for transmitting data or computing power for data analysis, but rather because there is no agreed global exchange system with standards and accepted processes for the collection, storage, access, and preservation of research data, with trained data professionals to support this exchange.

In academic publishing, peer review and citation have long been recognised as mechanisms for endorsing the trustworthiness of scientific knowledge published in journals and books, incentivising researchers to contribute their results, and helping in the discovery and exchange of these research outputs. Trustworthy research data will only become widely available if similar principles are applied in data publishing.

Recently there has been an encouraging increase in the awareness and practice of referencing and citing research data, true to the predicted emergence of the '4th paradigm', Jim Gray's vision of "data-intensive scientific discovery", in Hey et al. (2009). Since its launch in 2009, the DataCite consortium has assigned some 2 million Digital Object Identifiers (DOIs) to help make research data discoverable and citable.[1] In 2011, Elsevier launched an initiative to link papers to the underlying data sets in many different repositories and databases (Aalbersberg and Kähler, 2011). In June 2012, the Association of Scientific, Technical and Medical publishers signed a joint statement with DataCite to encourage publishers and data centres to link papers and underlying data.[2]

In 2012, Thomson Reuters launched Data Citation Index, a new commercial service to assist in the tracking of data citations in the literature.[3] In 2013 the US National Science Foundation (NSF) the European Commission and the Australian Government have launched the Research Data Alliance (RDA) as another global platform for stakeholders to accelerate international data-driven innovation and discovery by facilitating research data sharing and exchange, use and re-use, standards harmonisation, and discoverability.[4]

### *1.2 Identifying authors and other knowledge contributors*

It is increasingly common that a given scholarly work – traditionally a journal paper or monograph, but more recently a range of other digital content published online – is unambiguously identified via its DOI, or via some other identifier scheme. However, determining which person or people contributed to a given work remains difficult. This is because contributors are identified in bibliographic records by name only, i.e., the ‘author’ or ‘creator’ metadata fields typically hold a simple text string rather than one or more unique person identifiers.

There is no guarantee of uniqueness for a person’s name; many names are shared by more than one person. Individuals can change their name during their lifetime. In addition, the same name can be expressed differently from one work to another due to variable use of initials, surname, or hyphens. Further complicating matters, a name can have different spellings after transliteration from Chinese, Russian, Irish, Scandinavian or other languages to Roman characters, as pointed out by Qiu (2008). It follows that names are unsuitable as globally unique identifiers.

Governments have addressed this problem by establishing national ID schemes to facilitate taxing and for a myriad other administrative purposes. However, national ID schemes are not used in all countries where scholarly research takes place, and only a small number of countries (including Norway and Iceland, Watson (2010)) use national IDs in widely accessible databases. Even so, as Adams (2013) says, research is by nature international in scope, with scholars frequently working at several institutions in different countries during their professional career, and so national ID schemes cannot solve more than a small slice of the problem.

In the context of the global scholarly literature corpus, an increasing population of contributors associated with a growing number and diversity of published works steadily expands the scale of the name ambiguity problem. For example, given that, as defended by Dogan et al. (2009), search by author name is one of the most common ways to query bibliographic databases (such as Web of Science, Scopus, or PubMed), name ambiguity creates obvious problems in navigating the scientific literature and understanding contributor networks. Thus name ambiguity is a fundamental obstacle to accurately attributing research products to the individuals who created or otherwise contributed to them.

### *1.3 Interoperability challenges: connecting contributors and data*

Many of these identification challenges have been addressed to some degree. For example, in some scientific disciplines data identification and data citation has reached a stage of high maturity and can be considered part of those communities’ norms. Also, the ORCID initiative (see below) is emerging as a global solution to the contributor identification problem.

However, lack of interoperability between identification systems for data on one hand and for contributors on the other remains a major hurdle. A formal data citation via a persistent identifier, for example, creates a link to the data centre where the cited dataset is published. But if the data creators are only referenced by name in the metadata describing the dataset, for the reasons outlined above it is frequently not possible to create a reliable link from the creators to the dataset. Therefore, it becomes difficult or impossible to associate creators with (for example) measures of impact and other downstream tracking of data use and reuse.

The ODIN project focuses on three threats or "items of unfinished business" emanating from lack of recognition of the need for robust ways of identifying contributors and their data:

- *Inability to follow interconnections between datasets and contributors as a method of data discovery*. It is currently impossible to guarantee direct access to every dataset that is used in a given piece of research as published in a journal paper. This is the case even when the data are made available by the authors, the publishers, or stored in a data repository. As the published literature continues to be the primary source of inspiration for new research, this situation inhibits re-use, verification of research, and detection of scientific fraud. Further, it is currently impossible to guarantee that potential re-users will identify, understand and comply with the conditions of re-use of specific datasets. While in some domains the concept of Open Data is crucial, there are often legal, ethical or commercial constraints in other fields, such as privacy concerns over identifying data in medical or social science records. Without robust mechanisms for ensuring that terms and conditions of use and re-use are propagated within large-scale, automated data harvesting and data mining tools, entire fields of scientific knowledge are rendered unusable.

- *Inability to share and connect identifiers of contributors and authors between different user communities*. It is currently impossible to reliably connect an individual researcher, institution, region, funding agency or country uniquely to their journal papers, datasets, or other scholarly works. In addition to systemic issues with name ambiguity and incompatibilities between competing identifier protocols, current systems are either focused on a particular country, discipline (or sub-discipline) or a single university. This also has bearing on authentication systems in use in different communities. While some have common lineage and are inter-operable at the resolver level, each relies on internally managed user identities and dataset identities, which makes it impossible to, for example, discover when two researchers co-own the same dataset, or one researcher owns two datasets. This limits the ability to realise an ecosystem in which multiple researchers asynchronously collaborate across multiple repositories.

- *Inability to uniquely identify datasets attributed to a particular contributor and contributors to a particular dataset*. It is currently impossible to ensure that a researcher will gain scientific credit for collecting, curating and publishing datasets. Treating datasets as independent citable records of science would establish a huge incentive for scientists to publish their datasets and share them with others. It will also assist institutions, funding agencies and policy makers in reliably determining which research outputs they have funded and monitoring their re-use and overall impact.

## 2 State-of-the-art

Two of the key premises of ODIN are that

- there already exists a diverse ecosystem of identifier systems in various stages of maturity, technical sophistication and scope (local, national, disciplinary, organisational etc.)
- that most if not all of these identifier systems will continue to find utility and be deployed within the environments where they first emerged.

A persistent identifier is characterised by

- a clear definition of the structure and syntax of the identifier itself
- a technical infrastructure for resolving the identifier.

A major aim of the ODIN project is to explore how, where practical, existing identifier systems can best interoperate.

Here we provide an overview of the two main relevant classes of identification infrastructure: for digital artefacts on the one hand and for people associated with these digital artefacts – authors, data creators, and others contributors – on the other.

### *2.1 Data as products of research*

Data are essential products resulting from research investigations (and other sources) and fundamental to basic scientific tenets such as reproducibility and transparency. As products, data should be labelled in ways that allow them to be reused. In fact, the new mode of data-intensive science makes the use of existing data a central asset of future science (Hey et al., 2009).

Data have always been the cornerstone of science; it is not possible to replicate experimental findings, perform observational research, or test assertions without data. Because data often have a longer lifecycle than the research projects that created them, understanding the role of data in the research lifecycle is vital.

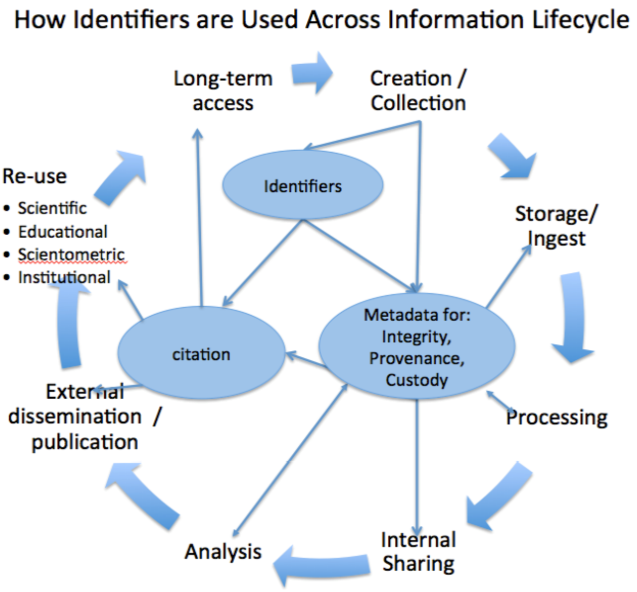
It is important to note that research lifecycles are as varied as the types of research performed, so generalisations are not always helpful. According to the current paradigm (Figure 1), data are integral to several steps in the research lifecycle: running the experiment, creating and collecting research results, and analysis. Ideally, data also should be part of disseminating results; otherwise, the link to the data is broken and the provenance of the results is in question. Data citation provides this link.

Identification of resources through identifiers such as DOI names or Uniform Resource Names (URN) is a well-known solution to the long-term preservation of references. This approach is already widely used in data curation and traditional publications. Paskin (2004) argues that in the case of electronic access to research data, references provided by means of identifiers allow location of the desired resource in a similar way that is reliable and available over a long time.

A persistent identifier clearly identifies units of intellectual creations in a digital surrounding and supports administration of these units irrespective of form and granularity. It facilitates formal citation of the digital resource – in our case a scientific dataset – in scholarly papers, but also enables unambiguous identification of the dataset in a wide variety of data management applications. More importantly, identifiers allow cross-linkage of digital resources, including the linking of datasets to reference papers or source datasets from which they have been derived. Finally, since the provision of the dataset identifier is achieved through a registration mechanism, it allows specialised actors of data curation to keep track of the resource, index it in large catalogues allowing other researchers to find it and thereby dramatically improve the potential impact of a dataset publication.

All the above aspects have been identified by the scientific community as valuable and crucial for a better usage of scientific datasets, Klump et al. (2006).

**Figure 1** Actions on data and their reliance on metadata, data citation, and data management (see online version for colours)



*Source*: Adapted from Altman (2012)

## 2.2 *Identification systems for data*

There are many different persistent identifiers for data in use worldwide. Other than accession numbers, the most commonly used identifiers are URNs, ARKs and Handles/ DOIs (Table 1).

DOIs have emerged as the most widely used citation standard in the scholarly and professional publishing domain. DOI names are used by the European Commission through its publication agency, the Office of Publications of the European Union (OPOCE), and by several thousand scientific societies, publishers and companies worldwide through associations. The largest of these is the non-profit organisation CrossRef, with several thousand members in the publishing sector.

**Table 1** Identifications systems for data

| | *Exemplar identifier* | *Summary* |
|---|---|---|
| *URN*: uniform resource name | urn:isbn:0451450523 | Introduced in 1994, formalised in 1997 and is now an IETF standard. No central governance, no central resolving infrastructure. Used by major national libraries in Europe. ISBNs for books are part of the URN system. No license costs involved for assigning URNs. Registration agency needs to establish an assigning and resolving infrastructure. Initiative to harmonise URN registration in Europe by the PersID project[1] |
| *ARK*: archival resource key | ark:/13030/tf5p30086k | Introduced in 1995. Not a formal standard but all ARKs follow the same structure and workflows.[2] No central resolver – organisations can sign up to become Name Assigning Authority Numbers (NAANs) and run their own resolution infrastructure for ARKs. System is run by the California Digital Library with dozens of NAANs worldwide through a combined ARK/DOI infrastructure EZID[3] |
| *Handle* | hdl:2381/12775 | Non-commercial decentralised identifier resolution system, established in 1995. Operated by the Corporation for National Research Initiatives (CNRI). Used by many higher-level systems, e.g., DOI. Different initiatives use commercial handle licenses to establish local handle system, such as the European Persistent Identifier Consortium (EPIC).[4] Many existing content management systems, including institutional repositories, currently operate their own local handle system |
| *DOI*: digital object identifier | doi:10.1186/2041-1480-3-9 | Combines a metadata model with the Handle system as the resolution infrastructure (i.e., DOIs are handles). First introduced in 1998 with the funding of the International DOI foundation (IDF). Became official ISO standard in 2012 (ISO 26324). The DOI system is built upon CNRI Handles. A standard metadata kernel is defined for every DOI name. Assigning DOI names involves the payment of a license fee but their resolution is free. DOI Registration agencies are responsible for assigning identifiers. The DOI system itself is maintained and advanced by the IDF, itself controlled by its registration agency members |

[1] http://www.persid.org

[2] https://confluence.ucop.edu/display/Curation/ARK

[3] http://n2t.net/ezid/

[4] http://www.pidconsortium.eu

However, while the interoperability and long-term preservation of linkage in scientific paper publication has been largely achieved through DOI over the last 10–12 years, dataset publication has not reached a similar maturity level. The issue of access to datasets has grown more and more important in the different European research areas. The Digital Curation Center in the UK,[5] for example, was established in 2007, but serves only as an advisory centre and does not itself provide storage of or access to datasets, nor does it issue data identifiers. Another attempt was started by the Alliance for Permanent Access, which aims to develop a shared vision and framework for a sustainable organisational infrastructure for permanent access to scientific information.[6] None of these approaches has yet established a workflow or a functional infrastructure for data registration.

### *2.3 DataCite and data DOIs*

DataCite is an international consortium of 17 libraries and information institutions worldwide, led by the German National Library of Science and Technology (TIB). DataCite was founded on December 2009 as a global DOI registration Agency for Research Data[7] and has registered over 1.7 million data sets with DOI names so far. DataCite's use of the DOI system for registration allows scientists, data centres and publishers to use the same syntax and technical infrastructure for the referencing of datasets that are already established for the referencing of papers. For example, the following dataset:

> Storz, D *et al.* (2009): *Planktic Foraminiferal Flux and Faunal Composition of Sediment Trap L1_K276 in the Northeastern Atlantic*, PANGAEA data repository for earth and environmental science. http://dx.doi.org/10.1594/PANGAEA.724325

is used and cited in this paper:

> Storz, D., Schulz, H., Waniek, J.J., Schulz-Bull, D. and Kucera, M. (2009): *Seasonal and Interannual Variability of the Planktic Foraminiferal Flux in the Vicinity of the Azores Current*, Deep-Sea Research Part I-Oceanographic Research Papers, Vol. 56, No. 1, pp.107–124, http://dx.doi.org/10.1016/j.dsr.2008.08.009

A key reason why DOIs are suitable for datasets and data citation is that the DOI system is already widely accepted and understood by the publishing community and by academics more generally. However, DOIs have certain disadvantages in scenarios involving huge amounts of data, particularly for datasets that are dynamic, as by definition the content should not be altered once a DOI is assigned to it. For datasets still in the production cycle, DataCite suggest the use of other existing identifier systems, many of which have been established by DataCite members. For example, handles are widely used by the ANDS to identify data sets, whilst the California Digital Library uses ARKs for the same use cases. TIB assigns URNs in addition to DOIs for datasets in local German repositories.

To enable relations between different objects that have identifiers, DataCite has defined a set of metadata relations between datasets and other digital objects (e.g., isPartOf, isCitedBy, isSupplementTo, isCompiledBy, isVariantFormOf, isOriginalFormOf, isContinuedBy, etc.).

For additional services the exact type of dataset identifier is not the most important part, as the service expects the same outcome of any operation it performs with any data identifier. Therefore, it is crucial that the identifier communities harmonise their workflows, structures and metadata models to provide a seamless interaction between services based on different identifiers. This is already happening within DataCite with respect to the ARK, handle, and DOI communities, and also via cooperation between

DataCite and other communities and organisations, including the Knowledge Exchange initiative and its Den Haag manifesto.[8]

## 2.4 Identification systems for authors and other contributors

On the contributor side, every single actor from publishers to libraries, from repositories to large-scale participative infrastructures, has, on some level, its own contributor identifier 'layer'. There is huge diversity in functionality and technical sophistication of these existing infrastructures, sometimes conveniently grouped together under the umbrella term "contributor identifier system". Common to all is information relating to people which pertains in some way to their research related activities – as authors/creators of scholarly content, as repository submitters, data curators, and so on. This implies a person identifier of some kind, which may be anything from a purely internal database or data file record reference with a local scope (e.g., a legacy library information system), to a globally unique, public, user-facing identifier with utility in cross-system integration such as ORCID.

The contributor identifier systems that have emerged in recent years in the scholarly domain range from nation/discipline/organisation-specific initiatives to services with a global scope (see Table 2). The former category includes a number of very successful services with deep adoption in the respective communities, some being tightly integrated into services or workflows that researchers use routinely, such as the RePEc Author Service,[9] a set of bibliographic indexes and related services for economics. In the latter category we find initiatives such as AuthorClaim (extension of the RePEc Author Service) and Thomson Reuter's ResearcherID, which have seen limited uptake in the broader research community for a variety of reasons, including lack of promotion by service providers and (in the case of ResearcherID) distrust in a service operated by a for-profit company.

## 2.5 ORCID as an identifier hub

The complexity of the problem space, and the fact that many of the problems in the scholarly domain need to be addressed globally rather than by discipline and/or locally, led to the ORCID initiative (http://orcid.org) which was started in 2009. As an organisation, ORCID's scope is truly international, transcends disciplinary boundaries, and has commitment from as varied a set of stakeholders as universities (Harvard, MIT, Cornell, Cambridge), corporations (Elsevier, Thomson Reuters, Wiley, Avedas), scholarly societies (APS, ACM, MLA, AGU), funders (Wellcome Trust, NIH, DOE, FDA, JSTA), data repositories (ANDS, CAS Library, Figshare, INSPIRE, arXiv) with over 100 integrations and over 1 million registrants worldwide in two years since launch of the ORCID Registry in October 2012.

The ORCID infrastructure is not intended to supplant all other contributor identifier systems, but rather to interoperate with and connect to these and other systems, including data registries. In particular, ORCID can serve as a kind of 'switchboard' or unifying integration point for incorporating contributor identifiers into a wide variety of research workflows, hitherto an impractical proposition due to the patchwork landscape of existing systems. Interoperability is possible in the ORCID framework through relations linking elements of ORCID records relating to the same contributor found in different systems or self-claimed (e.g., sameAs, submittedBy, and claimedBy).

**Table 2** Contributor identifier systems overview

| *Organisation* | *Type* | *Characteristics* | *Disciplines* | *Countries* | *Year started* |
|---|---|---|---|---|---|
| Open library society | Nonprofit | Integrates with databases for institutions (ARIW) and publications (3lib.org). RePEc Author Service extended as AuthorClaim in 2008 | All, currently mostly economics | All | 1999 |
| National council for scientific and technological development (CNPq) | Government | Part of several databases covering many scholarly activities. Mandatory for all Brasilian researchers since 2002 | All | Brazil | 1999 |
| Online computer library centre (OCLC) and 15 national libraries | Nonprofit | Integrates name authority records from several national libraries. Also contains other creators of creative content | All | Several | 2003 |
| Royal netherlands academy of arts and sciences (KNAW) | Government | Part of a database for publications, datasets and research projects | Part of a database for publications, datasets and research projects | Netherlands | 2004 |
| Cornell university library | Academic | Part of e-print archive (arXiv) | Physics, mathematics, computer science and related disciplines | All | 2005 |
| Elsevier | Commercial | Integrates with bibliographic database (Scopus) | All | All | 2006 |
| Mimas, British library | Academic | Identifiers for researchers and institutions | All | UK | 2007 |
| Thomson reuters | Commercial | Integrates with bibliographic database (Web of Science) | All | All | 2008 |
| ORCID | Nonprofit | Integrates with bibliographic databases and other author identifier systems | All | All | 2009 |

**Table 2** Contributor identifier systems overview (continued)

| *Organisation* | *Type* | *Characteristics* | *Disciplines* | *Countries* | *Year started* |
|---|---|---|---|---|---|
| National library of medicine (NLM) | Government | Part of several biomedical databases for publications and datasets (NCBI) | Life sciences | All | 2010 |
| CERN, DESY, Fermilab and SLAC | Academic | Digital library, integrated with preprint archives, journals, data centres and other information resources | High-energy physics | All | 2010 |
| ISNI international agency (ISNI-IA) | Nonprofit | Broad scope, partial overlap with ORCID. Authors and other persons or non-person entities relating to creative works, including companies and fictional characters | All | All | 2009 |

The main beneficiaries of the ORCID ID service are:

- *Researchers*, who can register for free to obtain a persistent, globally-unique person identifier that provides value to them in the form of reduced data entry and improved discoverability. ORCID identifiers are being integrated into a range of scholarly communication workflows including manuscript submission, grant application, dataset deposition and other contexts such as impact metrics.
- *ORCID members* and other organisations in the research and scholarly domain, who can extend their systems to connect to and integrate with the centralised service. This enables them to embed ORCID identifiers in their workflows, reduce problems of duplicate records, connect internal systems, and synchronise with external data enabling them to solve problems and create new opportunities.

A common feature of the systems listed in Table 2 is that they ‘sit below’ ORCID in the identifier system hierarchy. INSPIRE (the High Energy Physics (HEP) information system, http://inspirehep.net) is a case in point. Launched in 2010 and operated by CERN, DESY, Fermilab, and SLAC, INSPIRE is a digital library for the high-energy physics (HEP) community, linked with preprint archives, journals, large-scale data centres, institutional repositories and other important digital HEP resources.

In addition to common features such as searching across content of both the local system and linked remote repositories, INSPIRE performs author name disambiguation to create author profiles. Authors can register in the INSPIRE system, claim their author profiles, get their personalised author page, and claim their published papers.

Functionality provided by INSPIRE is very much tailored specifically to the HEP community, such as integration with large-scale data centres and with arXiv,[10] and provides value to researchers on top of the core mission of ORCID. That said, the international and interdisciplinary reach of ORCID provides a structure to link the HEP community into the worldwide scientific community. In addition, ORCID identifiers are becoming part of the metadata on research datasets, publications, and grants (etc.) and can be used to manage links between researchers and research. Linking ORCID identifiers to existing author identifier systems such as INSPIRE allows these services and their users to keep records up to date with minimal user intervention.

### *2.6 ORCID and ISNI*

Apart from ORCID itself, the exception from the characterisation above is the International Standard Name Identifier (ISNI: http://www.isni.org). This initiative, created around the same time as ORCID, aims to uniquely identify public entities across multiple fields of creative activity. Public entities can be people (including researchers), fictional characters, companies and institutions. Like ORCID, ISNI aims to connect numerous existing name authority systems by creating a 'meta-identifier' to facilitate integration.

Despite some overlap in scope, the two organisations have different missions, different ways of operating and different strategic approaches and priorities. For example, ISNI's approach is to 'seed' its registry with data from existing databases and collections (including databases built by national libraries) and end users will interact with registration agencies rather than directly with the central ISNI system. By contrast, ORCID operates a self-claim registry – which entails engaging directly with researchers and respecting their privacy preference – and emphasises direct integration with publishers, universities, funders and other stakeholder organisations in research.

The various differences notwithstanding, the two organisations agree that the points of overlap require collaboration and they have recently made public statements to this effect.[11] First steps have also been taken to enable future interoperability. Identifiers issued by ORCID are formatted according to the ISNI ISO standard (ISO 27729[12]) and are chosen from a block of numbers set aside for them by ISNI in order to avoid having the same number assigned to different people. Further, ORCID is leveraging ISNI organisation identifiers to support linkages with researcher affiliations.[13]

## 3 Solving the interoperability problem: the ODIN conceptual model

The ODIN Consortium has developed a conceptual model for addressing identifier interoperability challenges. The approach we have taken for the development of this model entails building a model of interoperability that is *open*, *discipline-neutral,* and *inclusive*; building upon existing e-Infrastructures where possible; focusing on data citation and attribution, and suggesting proof of concept studies for first practical implementations of this model.

The ODIN model consists of three layers of increasing complexity:

- *the trusted identifier layer* – criteria for persistent identifiers for objects and people
- *the data citation virtuous circle* linking research data and their contributors via data centres, DataCite, and ORCID
- *common data services e-infrastructures* which provide linked persistent identifiers in the data services e-infrastructures for the European e-Infrastructure framework

### *3.1 The trusted identifier layer*

The term persistent identifier is commonly used to describe long-lasting identifiers to digital objects. While persistence is an important feature of digital identifiers used for data and people, the term does not fully capture the features required for digital identifiers in the common data services e-infrastructures layer. Moreover, it has become clear from ODIN's discussions with many stakeholders – and has also been one of the main conclusions of the DIGOIDUNA study (DIGIODUNA, 2012) – that the most reasonable strategy going forward is one that supports better collaboration of existing initiatives. Best practices for collaboration already exist and thus need not be reinvented by ODIN, and are captured by the Den Haag Manifesto (http://www.knowledge-exchange.info/Default.aspx?ID=462)[14] proposed by the 2011 Knowledge Exchange Persistent Object Identifier workshop.

In light of the above, the ODIN model introduces the term *trusted identifier* to refer to digital identifiers that are *unique, persistent, descriptive, interoperable* and *governed.* In practical terms, this means that trusted identifiers have the following characteristics (in part inspired by the Den Haag Manifesto):

- Are unique on a global scale, allowing large numbers of unique identifiers?
- Resolve as HTTP universal resource identifiers (URIs) with support for content negotiation, and these URIs should be persistent.
- Come with metadata that describe their most relevant properties, including a minimum set of common metadata elements. A search of metadata elements across all trusted identifiers of that service should be possible.
- Are interoperable with other identifiers through metadata elements that describe their relationship?
- Are issued and managed by an organisation that focuses on that goal as its primary mission, has a sustainable business model and a critical mass of member organisations that have agreed to common procedures and policies, has a governing body, and is committed to using open technologies?

Trusted identifiers and Linked Open Data

Substantial momentum is gathering around Linked Data based approaches in the scholarly communication domain. The Linked Data[15] paradigm is based on a small set of core principles originally proposed by Tim Berners-Lee in 2006[16] (Berners-Lee et al., 2001) (Shadbolt et al., 2006) as a pragmatic implementation of the Semantic Web ideals. Those principles have since been expanded and built upon by the broader Linked Open Data (LOD) community. In a nutshell, LOD involves publishing data in such a way that makes it useful and enables interlinking by using HTTP-resolvable URIs to identify things (both physical entities and concepts). In turn, those URIs provide useful information in a machine-readable structured form following the standard RDF data model,[17] using shared vocabularies of properties to describe those things.

The main hurdle to widespread use and utility of Linked Data in the scholarly domain is lack of shared ontologies of properties (aka terms) for describing things or, more commonly, insufficient use of existing ontologies. Shared ontologies – structured, controlled vocabularies of terms for concepts, their definitions and well-defined relationships between them – enable joining and querying of datasets based on properties with the same or similar meaning, such as the type of a published work, or type of external contributor identifier. In other words, ontologies serve as a ‘semantic layer’ to convey the meaning of data exchanged between systems and enable the use of sophisticated informatics technologies to drive query tools and knowledge-discovery applications.

DataCite (in close collaboration with CrossRef) has for about two years returned DOI metadata as RDF in response to queries to their API. ORCID recently implemented experimental Linked Data support in its public API. Other key initiatives relevant to ODIN are using or adopting LOD approaches:

- The Common European Research Information Format (CERIF)[18] specifies a comprehensive data model and relational database environment for research information. Originally developed in the 1980s and developed since early 2009’s by the non-profit organisation EuroCRIS, CERIF’s focus has historically been on describing so-called current research information systems (CRIS) in academic institutions. Over time CERIF has evolved into an industry standard and is supported by most commercial CRIS software vendors. EuroCRIS is currently working on ways to expose CERIF datasets as Linked Data, via two complementary resources: the CERIF Ontology[19] and the CERIF Semantic Vocabulary.[20]
- The Consortia Advancing Standards in Research Administration Information (CASRAI)[21] is a closely related initiative. The CASRAI data dictionary (not published as a formal machine-readable ontology) and so-called community profiles closely map to the CERIF specification and can be implemented in CERIF.[22]
- The Semantic Publishing Project[23] is a more recent, fully Semantic Web based initiative which has created a family of small, complementary ontologies for describing bibliographic records and research information. SCoRO,[24] FaBiO[25] and FRAPO[26] ontologies are most relevant to ODIN. Crucially, ScoRO and FRAPO are based on, and are compatible with, the CERIF ontologies and so are complementary to the aforementioned standards.

Given the overall agreement on Linked Data principles in both the persistent identifier and semantic web communities, ODIN currently sees no need for further discussions among the different stakeholder groups. Rather, we recommend the practical implementation of these principles. Specifically, ODIN will:

- coordinate a minimum set of common schema elements between DataCite and ORCID
- explore options for adopting CERIF-based metadata scheme elements where appropriate
- work on interoperability based on common metadata schemas with other stakeholders

### *3.2 The data citation virtuous circle*

Data citation enables easy reuse and verification of data, allows the impact of data to be tracked, and creates a scholarly structure that recognises and rewards data producers.[27] An essential part of data citation is the linking of persistent identifiers for the data with persistent identifiers for contributors. ODIN therefore extends the data citation model to include these contributor identifiers.

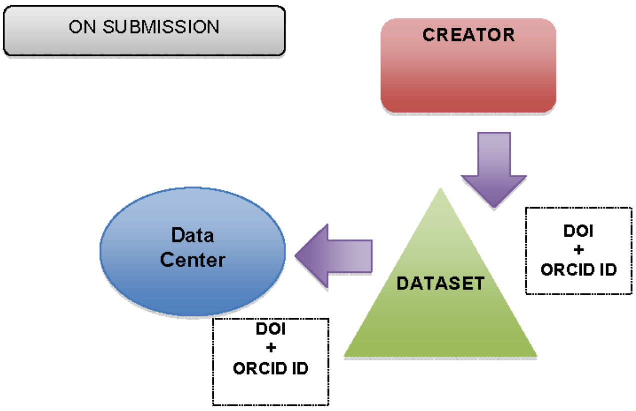
It is important to look at some key aspects of the workflow. First, the information that enables data citations is created when researchers submit datasets to data centres. It is therefore crucial that data centres work closely with both DataCite and ORCID so that the relevant identifiers for data and contributors can be created/retrieved and added to the dataset. Second, many datasets and persistent identifiers for data have been created to date. To get credit for these existing data publications, researchers need to be able to claim them; that is, to link the publications to their contributor identifier.

These two main scenarios – submitting data and claiming published data – involve the same actors: the researcher, the data centre, DataCite, and ORCID. The three organisations are connected to each other in what we call the *data citation virtuous circle*: information flows from the data centre to the DataCite Metadata Store (MDS), from the DataCite MDS to the ORCID Registry, and finally from the ORCID Registry back to the data centre, with information getting enriched in every step of the cycle. The flow is of course not strictly unidirectional, as lookup services are needed at every step, e.g., from the data centre to ORCID when a dataset is submitted.

### 3.3 Scenario A: submission of datasets to the data centre

When a researcher deposits a dataset with a data centre, his or her ORCID identifier – as well as the identifiers of co-contributors – should be linked to the dataset (Figure 2). A common related scenario is the submission of one or more datasets at the same time as a primary research paper, as is standard practice for datasets submitted to the Dryad repository.[28]

**Figure 2** Submission workflow (see online version for colours)



Other identifiers for data and/or authors can also be included in the metadata. Datasets are often part of larger collections, can exist in different versions, and frequently have associated documentation and specialised software to analyse them. The DataCite Metadata Schema (DataCite Metadata Working Group, 2011) facilitates the capture of this information in the dataset metadata. Information about the kind of contribution may also be added upon submission to the data centre.

The metadata are forwarded from the DataCite Metadata Store to the ORCID Registry. This step should happen automatically for all datasets that include ORCIDs in their metadata. For this step to work we need protocols for technical interoperability, and we need a minimal set of metadata that are standardised between ORCID and DataCite. Contributor roles currently supported by DataCite are listed in Table 3.

**Table 3** DataCite resource and contributor types

| ResourceTypeGeneral | ContributorType |
|---|---|
| Collection | ContactPerson |
| Dataset | DataCollector |
| Event | DataManager |
| Film | Distributor |
| Image | Editor |
| InteractiveResource | Funder |
| Model | HostingInstitution |
| PhysicalObject | Producer |
| Service | ProjectLeader |
| Software | ProjectMember |
| Sound | RegistrationAgency |
| Text | RegistrationAuthority |
| | RelatedPerson |
| | Researcher |
| | RightsHolder |
| | Sponsor |
| | Supervisor |
| | WorkPackageLeader |

*Source*: DataCite Metadata Working Group. DataCite Metadata Schema for the Publication and Citation of Research Data. DataCite; 2011: http://dx.doi.org/10.5438/0005

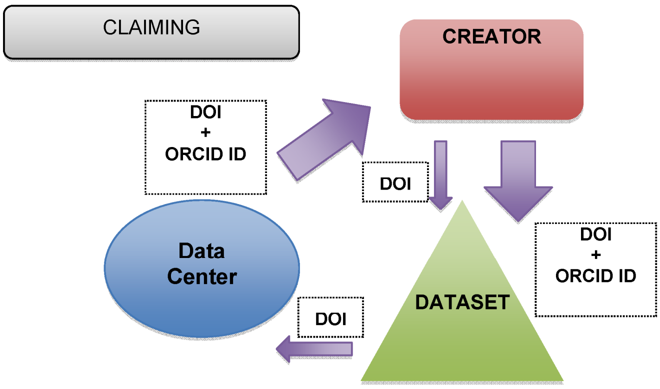
Integration of datasets and author identifiers in the data centre is the biggest challenge for interoperability, because it requires a different technical implementation for every data centre. To address this issue, ODIN is working on several proof-of-concept technical implementations.

### *3.4 Scenario B: claiming of already published datasets in the ORCID registry*

For datasets that are already published (over 1.7 million DataCite data DOIs have been assigned to date), we need to retrospectively link them to their authors. We expect this claiming to be necessary until DataCite DOIs and ORCID identifiers are routinely included with datasets intended for citation (Figure 3).

ORCID provides services to allow researchers to retrospectively claim their works, and to embed their identifier as they publish new works. Exemplifying how self-claiming can work in practice, ORCID has released a service[29] that enables researchers to claim publications in the DataCite Metadata Store via the ORCID registry. Similar to the automated transfer of metadata from DataCite to ORCID, this requires harmonisation of the metadata formats between DataCite and ORCID. A similar approach being adopted by ANDS enables linking datasets in Research Data Australia (RDA)[30] to ORCID identifiers. The new service is currently in the development phase and was made available as part of release R11 of RDA in late 2013.

**Figure 3** Claiming workflow (see online version for colours)



Claiming can also be performed by an ORCID member institution on behalf of the researcher. The workflow and tools required are much the same as the self-claim case, though an additional step is needed to first establish that the institution has the authority to make these claims. The institution plays an important role in supporting its researchers in documenting their research outputs, and this support is often but not exclusively provided by libraries. The first universities have started the technical integration with the ORCID service, and they can assist their researchers in claiming their datasets and other research outputs.

To disseminate the link between dataset and contributor from the ORCID Registry, the claims need to be made available for reuse by other parties. This includes (but is not limited to) reuse by bibliographic databases. ORCID provides an open API for this purpose. All profile data in ORCID are marked by the contributor as publicly available are free to be reused under a CC0 waiver,[31] explicitly placing the data in the public domain, as per the ORCID principles.[32] The metadata also include provenance information indicating whether the claims are self-claims made by researchers, and/or claims verified by the data centre.

The link between author and data also needs to be distributed back to the data centre that created the DOI for the dataset. The data centre will validate claims made by researchers and subsequently send the updated metadata to DataCite – this is the only way the DataCite Metadata Store can be updated.

In the final step, DataCite updates the ORCID Registry. The dataset is already linked to the ORCID identifier, but this claim is now verified by the data centre which adds an extra level of trust.

### *3.5 Common data services e-infrastructures*

The previous section focused on the initial steps of linking data and contributors. Equally important is the next step, which takes advantage of the interoperability layer to navigate across data and contributors. This will allow the scholarly community to expand their view from a paper-centric to an artefact-centric and contributor-centric perspective.

In the ODIN project we are focusing on services enabling linkages between data and publications, and to support impact assessment. By following the principles for trusted identifiers described above, we are creating a persistent identifier layer that can facilitate interoperation with community-supported e-Infrastructures, data curation, data preservation, authentication/authorisation and other e-Infrastructures services. Importantly, the persistent identifier layer is independent and not tied to related, but separate, e-Infrastructure, such as authentication or infrastructure for open access. This will decrease the dependencies on other infrastructure, in turn making the persistent identifier layer easier to maintain and less likely to fail. This will also increase the potential for interoperability, for example with infrastructures outside of Europe.

- *Linking data and publications*. One important example of interoperability beyond data and people is publications. Datasets are frequently associated with publications citing them, and we want to navigate from datasets to publications via links based on DOIs and other trusted identifiers for these different types of research outputs. Although a lot of these connections between data and publications already exist, an interoperability layer will vastly facilitate making these connections and the connections from publications to authors.
- *Impact assessment*. Scientific impact today is often demonstrated by citations, and we have scientific infrastructure to track citations to scholarly papers. The persistent identifier layer will extend this infrastructure to citations for data. This will make it much easier – and more accurate – to track the impact of individual researchers and all their research contributions.

## 4 Conclusions and next steps

Unique identification of both research outputs and contributors remains a challenge in many fields of knowledge. However, the maturity of initiatives such as DataCite and ORCID invite the possibility of taking a step forward and defining an interoperability framework. From a conceptual point of view, this task is feasible, not only for new research outputs but also for existing outputs. The interoperability layer built by DataCite and ORCID enables community-supported e-Infrastructures to share discipline-neutral, interoperable, open and persistent identifiers for data and contributors.

## Acknowledgements

The ODIN project received funding from the European Union's Seventh Framework Program for research, technological development and demonstration under Grant Agreement No. 312788.

## References

Aalbersberg, I-J. and Kähler, O. (2011) 'Supporting science through the interoperability of data and articles', *DLib Magazine*, Vol. 17, Nos. 1–2, January–February, http://dx.doi.org/10.1045/january2011-aalbersberg

Adams, J. (2013) 'Collaborations: the fourth age of research', *Nature*, Vol. 497, pp.557–560 http://dx.doi.org/10.1038/497557a

Altman, M. (2012) *Needs for Data Management & Citation Throughout the Information Lifecycle. Prepared for NISO Forum: Tracking it Back to the Source: Managing and Citing Research Data September 2012 Needs for Data Management &Citation Throughout the Information Lifecycle*, http://www.slideshare.net/drmaltman/needs-for-data-management-citation-throughout-the-information-lifecycle

Berners-Lee, T., Hendler, J. and Lassila, O. (2001) 'The semantic web – a new form of web content that is meaningful to computers will unleash a revolution of new possibilities', *Scientific American*, Vol. 284, pp.34–43.

DataCite Metadata Working Group (2011) *DataCite Metadata Schema for the Publication and Citation of Research Data [Internet]*, DataCite e.V.; 2011: http://dx.doi.org/10.5438/0006

DIGIODUNA (2012) *Digital Object Identifiers and Unique Authors Identifiers to Enable Services for Data Quality Assessment, Provenance, and Access*, 2012: http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/digoiduna.pdf

Dogan, R.I., Murray, G.C., Névéol, A. and Lu, Z. (2009) *Understanding PubMed User Search Behavior through Log Analysis*, Database bap018 (2009), http://dx.doi.org/10.1093/database/bap018

Hey, T., Tansley, S. and Tolle, K. (Eds.) (2009) *The Fourth Paradigm: Data Intensive Scientific Discovery*, Microsoft Research, Redmond, WA, Retrieved from http://research.microsoft.com/en-us/collaboration/fourthparadigm/

Klump, J., Bertelmann, R., Brase, J., Diepenbroek, M., Grobe, H., Höck, H., Lautenschlager, M., Schindler, U., Sens, I. and Wöchter, J. (2006) 'Data publication in the open access initiative', *Data Science Journal*, Vol. 5, pp.79–83.

Paskin, N. (2004) 'Digital object identifiers for scientific data sets', *19th International CODATA Conference*, Berlin, Germany.

Qiu, J. (2008) 'Scientific publishing: Identity crisis', *Nature News*, Vol. 451, pp.766.

Shadbolt, N., Lee, T.B. and Hall, W. (2006) 'The semantic web revisited', *IEEE Intelligent Systems*, Vol. 21, pp.96–101.

Watson, I. (2010) 'A short history of national identification numbering in Iceland', *Bifröst Journal of Social Science*, Vol. 4, pp.51–89, http://bjss.bifrost.is/index.php/bjss/article/view/63

## Notes

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

[2] http://www.datacite.org/node/65

[3] http://wokinfo.com/products_tools/multidisciplinary/dci/

[4] http://rd-alliance.org

[5] http://www.dcc.ac.uk/about

[6] http://www.alliancepermanentaccess.eu

[7] http://www.datacite.org

[8] http://www.knowledge-exchange.info/Default.aspx?ID=462

[9] https://authors.repec.org

[10] http://arxiv.org

[11] http://orcid.org/blog/2013/04/22/orcid-and-isni-issue-joint-statement-interoperation-april-2013

[12] http://www.isni.org/content/iso-publishes-isni-standard-iso-277292012

[13] http://orcid.org/blog/2013/06/27/orcid-plans-launch-affiliation-module-using-isni-and-ringgold-organisation

[14] Den Haag Persistent Object Identifier – Linked Open Data Manifesto: http://www.knowledge-exchange.info/Default.aspx?ID=462

[15] http://en.wikipedia.org/wiki/Linked_Data

[16] http://www.w3.org/DesignIssues/LinkedData.html

[17] RDF Primer: http://www.w3.org/TR/rdf-primer/

[18] http://www.eurocris.org/Index.php?page=CERIFintroduction&t=1

[19] http://www.eurocris.org/ontologies/cerif/1.3

[20] http://www.eurocris.org/ontologies/semcerif/1.3

[21] http://casrai.org

[22] http://www.cerifsupport.org/2012/12/17/cerif-casrai-profiling/

[23] http://semanticpublishing.wordpress.com

[24] SCoRO, the Scholarly Contributions and Roles Ontology: http://purl.org/spar/scoro/

[25] FaBiO, the FRBR-aligned Bibliographic Ontology: http://purl.org/spar/fabio/

[26] FRAPO, the Funding, Research Administration and Projects Ontology: http://purl.org/cerif/frapo/

[27] http://www.datacite.org/whycitedata

[28] http://datadryad.org

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

[30] http://researchdata.ands.org.au

[31] http://creativecommons.org/about/cc0

[32] http://orcid.org/about/what-is-orcid/our-principles

## Website

Den Haag Persistent Object Identifier – Linked Open Data Manifesto: http://www.knowledge-exchange.info/Default.aspx?ID=462