Grant agreement no. 312788

ODIN

ORCID AND DATACITE INTEROPERABILITY NETWORK

www.odin-project.eu

D4.1 Conceptual model of interoperability

WP4 – Interoperability

V1.0

Final

Abstract: A conceptual model to solve the interoperability between different identifiers for data and people is proposed. The model describes key criteria for persistent identifiers and metadata describing the relationship between different identifier schemes. We describe how this conceptual model fits into the Linked Open Data model and look at several key relevant workflows.

SEVENTH FRAMEWORK PROGRAMME

Lead beneficiary: DataCite
Date: 03/08/2013
Nature: Report
Dissemination level: PU (Public)

Document Information

Grant Agreement no.	312788	**Acronym**	ODIN
Full title	ORCID and DataCite Interoperability Network		
Project URL	http://odin-project.eu		
Project Coordinator	John Kaye (BL) Address: The British Library 96 Euston Road, London NW1 2DB, United Kingdom Phone: +44 20 7412 7450 Email: john.kaye@bl.uk		

Deliverable	**Number**	4.1	**Title**	Conceptual model of interoperability	
Work package	**Number**	4	**Title**	Interoperability	

Document identifier	ODIN-WP4-Conceptual-Model-Interop-0001-1_0			
Delivery date	**Contractual**	July 2013	**Actual**	2nd August 2013
Status	Version 1_0		Final Draft □	
Nature	Report ☑ Prototype □ Demonstrator □ Other □			
Dissemination Level	☑ Public □ Restricted to other programme participants (including the Commission Services) □ Restricted to a specified group (including the Commission Services) □ Confidential, only for consortium members (including the Commission Services)			

Authors (Partner)	Martin Fenner (ORCID EU), Gudmundur A. Thorisson (ORCID EU), Sergio Ruiz (DataCite) and Jan Brase (DataCite)			
Responsible Author	Sergio Ruiz		**Email**	Sergio.Ruiz@datacite.org
	Partner	DataCite	**Phone**	

Document Status Sheet

Issue	Date	Comment	Author
0_1	23 Apr 2013	Initial draft	Sergio Ruiz (DataCite)
0_2	26 Apr 2013	Revised draft	Gudmundur A. Thorisson (ORCID EU)
0_3	30 Apr 2013	Revised draft	Martin Fenner (ORCID EU)
0_4	07 May 2013	Revised draft	Jan Brase (DataCite)
0_5	09 May 2013	First version ready for internal distribution	Sergio Ruiz (DataCite)
0_6	27 May 2013	1st round comments	Todd Vision (Dryad)
0_7	4 Jun 2013	Incorporating comments from external reviewer (David Shotton, Oxford University)	Gudmundur. A. Thorisson (ORCID EU) Martin Fenner (ORCID EU), Sergio Ruiz (DataCita), Jan Brase (DataCite)
1_0	02 Aug 2013	Final version including internal reviews (Laure Haak, ORCID and Amir Aryani, ANDS)	Sergio Ruiz

Document Change Record

Issue	Item	Reason for Change

TABLE OF CONTENTS

EXECUTIVE SUMMARY

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. At the same time, steps are being taken to identify the individuals who created or contributed to research outputs, and thus to address the so-called "author name problem". However, lack of interoperability between the different initiatives to identify datasets and contributors remains a major hurdle.

This deliverable presents an overview of the status of different persistent identifiers systems and initiatives, both for data and for contributors, and describes ODIN's conceptual model for tackling the associated interoperability challenges. In the overview of the state-of-the-art, special consideration is given to the infrastructures created and operated by ODIN partners DataCite and ORCID. In our conceptual model, we propose the concept of "trusted identifiers" and describe additional criteria that are required for persistent identifiers to become building blocks of a common data e-infrastructure. We discuss the importance of metadata schemas for research information such as CERIF, and how our conceptual model can be integrated with Linked Open Data. Different scenarios involving linking datasets with their creators are discussed, in particular concerning integrating them in the context of the Collaborative Data Infrastructure proposed by the High Level Expert group on Scientific Data (HLEG)[1].

[1] Riding the wave - How Europe can gain from the rising tide of scientific data. Final report of the High Level Expert Group on Scientific Data, October 2010: http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf

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-infrastructure. This project aims to support and stimulate the adoption of interoperable identifiers for researchers, research works and their outputs (publications and data). In addition, it facilitates information flow between research communities, leading to greater re-use of data and innovative exploitation of the existing knowledge.

ODIN will provide a roadmap which focuses on the integration and scalability of the DataCite and ORCID persistent identifier initiatives for tackling ODIN's four main challenges concerning research data: *accessibility*, *discovery*, *interoperability*, and *sustainability*. Work Package 4 focuses on one of these main challenges: interoperability between open identifiers for data and contributors in different infrastructures.

1.1. Identifying, discovering and citing data

Research data are arguably the world's greatest untapped resource. 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, and it can be the basis of economic systems that could lift millions more out of poverty. Research data have the capacity, but the world lacks the infrastructure to realize the benefits of that capacity. Useful free flow of data is currently not possible – not because of the absence of networks or computation, but 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, incentivizing researchers to contribute their results, and helping in the discovery and exchange of these research outputs. Trustworthy research data will only be widely available if similar principles are applied.

Persistent identifiers for research data. 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"[2]. 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[3]. In 2011, Elsevier launched an initiative to link articles to the underlying data sets in many different repositories and databases[4], in June 2012, the Association of

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

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

[4] Supporting Science through the Interoperability of Data and Articles", I-J. Aalbersberg, O. Kähler, DLib Magazine, January/February 2011 – Volume 17 Issue 1/2, http://dx.doi.org/10.1045/january2011-aalbersberg

Scientific, Technical and Medical publishers signed a joint statement with DataCite to encourage publishers and data centers to link articles and underlying data[5]. In 2012, Thomson Reuters launched a Data Citation Index to assist in the tracking of data citations in the literature[6]. In 2013 the NSF (National Science Foundation, US), 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 harmonization, and discoverability.[7]

1.2. Identifying authors and other knowledge contributors

It is increasingly common that a given scholarly work - traditionally a journal article 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 have historically been – and continue to be - 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. Moreover, some individuals change their name during their life, e.g., after marriage. Further complicating matters, the same name can be expressed differently from one work to another due to variable use of initials, surname, or hyphens. Further, a name can have different spellings after transliteration from Chinese, Russian, Irish, Scandinavian or other languages to Roman characters. It follows that names are unsuitable as globally unique identifiers. The governments addressed this problem by establishing national ID schemes which facilitate taxing and a myriad of other administrative purposes.

National ID schemes are not used in all countries where scholarly research takes place, however, and only a small number of countries including Norway and Iceland[8] use national IDs in widely accessible databases. Even so, research is by nature international in scope[9], with scholars frequently working at several institutions in different countries during their professional career, and so national ID schemes can not solve more than a small slice of the problem.

A big problem growing bigger. In the context of the global scholarly literature corpus, a growing population of contributors associated with a growing number and diversity of published works steadily increases the scale of the problem. Also, with increasing

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

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

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

[8] Watson, I. A short history of national identification numbering in Iceland. Bifröst Journal of Social Science 4 (2010): 51-89. http://bjss.bifrost.is/index.php/bjss/article/view/63

[9] Adams, J. Collaborations: The fourth age of research. Nature 497, 557–560 (2013). http://dx.doi.org/10.1038/497557a

participation by scholars from non-Western countries in recent years, the necessity of making author names fit into Western conventions and character sets has caused further confusion[10]. Given that search by author name is one of the most frequent ways to query bibliographic databases (such as Web of Science, Scopus, or PubMed[11]), name ambiguity creates obvious problems in navigating the scientific literature and understanding contributor networks. More generally, the so-called 'author name problem' is a fundamental obstacle to unambiguously attributing research products to the individuals who created or otherwise contributed to them.

1.3. Interoperability challenges - connecting contributors and data

As further discussed in the **State of the Art** section, many of the above 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 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 hand 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, 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 proposal highlighted three main threats or "items of unfinished business" emanating from lack of recognition of the need for robust ways of identifying contributors and their data in e-Science:

- Inability to follow interconnections between datasets and contributors as a method of data discovery.
- Inability to share and connect identifiers of contributors and authors between different user communities.
- Inability to uniquely identify datasets attributed to a particular contributor and contributors to a particular dataset.

Appendix B provides details for a series of scenarios identified by ODIN partners which further illustrate these points.

The next section provides an overview of current state-of-the-art technical solutions for identification of contributors and data, with a special focus on DataCite and Open Researcher & Contributor ID initiative (ORCID) infrastructures. The third section is the core of this deliverable and describes the conceptual model being developed in the

[10] Qiu, J. Scientific publishing: Identity crisis. Nature News 451, 766 (2008). http://dx.doi.org/10.1038/451766a

[11] Dogan, R.I. et al. Understanding PubMed user search behavior through log analysis. Database bap018 (2009). http://dx.doi.org/10.1093/database/bap018

ODIN project for addressing interoperability challenges. The final section provides conclusions and future outlook.

2. STATE OF THE ART

Two of the key premises of ODIN are that i) there already exists a diverse ecosystem of identifier systems in various stages of maturity, technical sophistication and scope (local, national, disciplinary, organizational etc.), and ii) 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 major aim of the project, therefore, is to explore how, where practical, these existing systems can best interoperate.

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

2.1. *Raison d'être* of persistent identifiers

A persistent identifier is characterized by: i) a clear definition of the structure and syntax of the identifier itself, and ii) a technical infrastructure for resolving the identifier.

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[12]. Data have always been the cornerstone of science, as 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 generalizations are not always helpful.

The current paradigm is that data are integral to the research steps “run experiment,” “create data,” “collect data,” and “analyze data.” Ideally, data also should be part of the “disseminate results” step; otherwise, the link from results back to the data is broken and the provenance of the results is in question. Data citation provides this link.

Figure 1 below shows a schematic diagram of the research lifecycle and how identifiers and other metadata are critical for these processes. It illustrates what data processes can be applied to data in the “run experiment,” “create data,” “collect data” and “analyze data” stages of the research lifecycle.

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

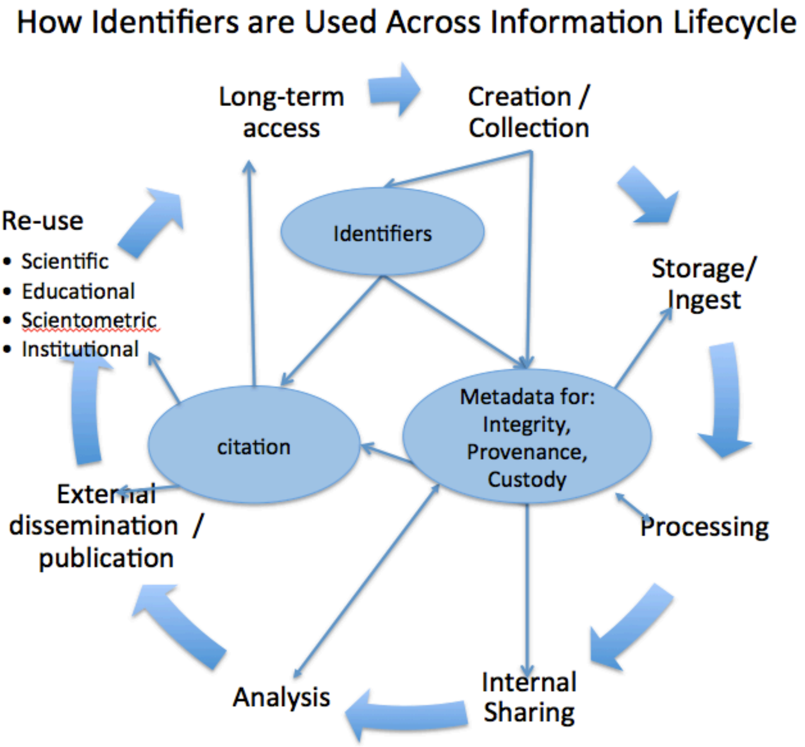


Fig. 1: Actions on data and their reliance on metadata, data citation, and data management (adapted from Altman, 2012).

Identification of resources through identifiers such as Digital Object Identifiers (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. For 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[13].

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 allows the citation of the digital resource (in our case a scientific dataset) in scholarly articles 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 articles 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 specialized 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

[13] Paskin, N. (2004) Digital Object Identifiers for scientific data sets. 19th International CODATA Conference, Berlin, Germany

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[14].

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 (see overview in Table 1).

	Exemplar identifier	Summary
URN - Uniform Resource Name	urn:isbn:0451450523	Introduced in 1994, formalized 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, but a URN registration agency needs to establish an assigning and a resolving infrastructure. The biggest initiative to harmonize URN registration in Europe is currently undertaken by the PersID project[15].
ARK - Archival Resource Key	ark:/13030/tf5p30086k	Introduced in 1995. Not a formal standard but all ARKs follow the same structure and workflows[16]. 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 [17]
Handle	hdl:2381/12775	Non-commercial decentralized identifier resolution system, established in 1995. Operated by CNRI. Used by many other higher-level systems, e.g. DOI. A non-commercial Handle system that is operated by the Corporation for National Research Initiatives (CNRI) Different initiatives use commercial handle licenses to establish local handle system, such as the European Persistent Identifier Consortium (EPIC)[18]. Many existing content management systems, including institutional repositories, currently operate their own local handle system.

[14] J. Klump et al. Data publication in the Open Access initiative. Data Science Journal, Volume 5 (2006). pp.79-83. http://dx.doi.org/10.2481/dsj.5.79

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

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

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

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

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. DOI Registration agencies are responsible for assigning identifiers. They each have their own commercial or non-commercial business model for supporting the associated costs. The DOI system itself is maintained and advanced by the IDF, itself controlled by its registration agency members. Using the Handle system, there is a central free worldwide resolving mechanism for DOI names. DOI names from any registration agency can be resolved worldwide in every handle server; DOIs therefore are self-sufficient and their resolution does not depend on a single agency. 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 is the preferred identifier system for DataCite.

Table 1: Identifications systems for data

Identification of scholarly works through DOI names

DOIs have emerged as the most widely used citation standard in the publication world. 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 which is the non-profit organization CrossRef, which has several thousand members in the publishing sector).

However, while the interoperability and long-term preservation of linkage in scientific article 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[19], 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 organizational infrastructure for permanent access to scientific information[20]. None of these approaches has however yet established a workflow or a functional infrastructure for data registration.

DataCite and data DOIs

DataCite is an international consortium of 17 libraries and information institutions

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

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

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[21] 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 centers and publishers to use the same syntax and technical infrastructure for the referencing of datasets that are already established for the referencing of articles. 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 article:

> Storz, David; Schulz, Hartmut; Waniek, Joanna J; Schulz-Bull, Detlef; Kucera, Michal (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, 56(1), 107-124,
> http://dx.doi.org/10.1016/j.dsr.2008.08.009

Interoperability of identifiers within the DataCite system

A key reason why DOIs are suitable for datasets and data citation is that the DOI system is 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. One reason for this is that, while there is no logical restriction to the number of DOIs that can be created, there is a cost associated with each DOI registered.

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 third party applications, data identifier interoperability is less of a technical challenge aand more an administrative challenge. For additional services the exact type of dataset identifier is not the most important part, as the service expects the same outcome of

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

any operation it performs with any data identifier. Therefore, it is crucial that the identifier communities harmonize their workflows, structures and metadata models to provide a seamless interaction between services based on different identifiers. This is already happening inside DataCite with respect to the ARK, handle, and DOI communities, but also in cooperation between DataCite and other communities and organizations, including Knowledge Exchange and its Den Haag manifesto[22] (discussed in more depth below).

2.3. 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”. What they all have in common 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 (ORCID, see below).

The contributor identifier systems that have emerged in recent years in the scholarly domain range from nation/discipline/organization-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[23], a set of bibliographic indexes and related services for economics.

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

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

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.

ORCID as an identifier hub

The complexity of the problem space, and the fact that many of the problems in the scholarly domain needed 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 organization, 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, KNODE, CAS Library, figshare, INSPIRE, arXiv), with over 70 members and over 200,000 registrants worldwide 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 as outlined above.

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).

The main customers of the ORCID ID service are:

- *Researchers*, who can register for free to obtain a persistent, globally-unique person identifier which 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 organizations in the scholarly domain, who can extend their systems to connect to and integrate with the centralized service. This enables them to embed ORCID identifiers in their workflows, reduce problems of duplicate records, connect internal systems, and synchronize with external data enabling them to solve problems and create new opportunities.

ORCID and local systems

A common feature of the systems listed in Table 2 is that they are situated closer to the end user and 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 centers, 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 personalized author page, and claim their published papers.

Whilst some of these features are shared with the central ORCID system, other functionality provided by INSPIRE is very much tailored specifically to the HEP community, such as integration with large-scale data centers and with arXiv[24] where >90% of HEP papers are freely available as full-text preprints. Due to such specialized functionality, INSPIRE, RePEc and other services similarly geared to specific communities/disciplines provide value to researchers that cannot easily be met by a generic service such as ORCID.

Technical functionality aside, it is also the case that community services are often operated by major organizations which are integral to their respective fields of research, such as CERN operating INSPIRE for the HEP community. This undoubtedly engenders a level of community trust unreachable by an organization such as ORCID, which is one step removed from researchers[25]. 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.

ORCID and ISNI

Apart from ORCID itself, the exception from the characterization 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

[24] http://arxiv.org

[25] The organizations which operate INSPIRE and other major HEP resources are represented on the ORCID Board of Directors. See http://orcid.org/about/team and http://orcid.org/about/what-is-orcid/governance

numerous existing name authority systems by creating a “meta-identifier” to facilitate integration.

Despite some overlap in scope, the two organizations 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 emphasizes direct integration with publishers, universities, funders and other stakeholder organizations in research.

The various differences notwithstanding, the two organizations agree that the points of overlap require collaboration and they have recently made public statements to this effect[26]. First steps have also been taken to enable future interoperability. Identifiers issued by ORCID are formatted according to the ISNI ISO standard (ISO 27729[27]) 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 will be leveraging ISNI organization identifiers to support linkages with researcher affiliations.[28]

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

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

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

3. SOLVING THE INTEROPERABILITY PROBLEM - THE ODIN CONCEPTUAL MODEL

The ODIN Consortium has developed a conceptual model for addressing the interoperability challenges outlined in section 1.3 and summarized below:

- Inability to follow interconnections between datasets and contributors as a method of data discovery
- Inability to share and connect identifiers of contributors and authors between different user communities
- Inability to uniquely identify datasets attributed to a particular contributor and contributors to a particular dataset

The approach we have taken for the development of this model entails the following:

- 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
- Suggesting proof of concept studies for first practical implementations of this model

This document focuses on the conceptual model, whereas a future report (D4.2 due at the end of the ODIN project) will discuss in detail a number of key workflows and specific technical interoperability issues.

3.1. Conceptual model - overview

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 centers, 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

3.2. 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 doesn't 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[29] – that the most reasonable strategy going

[29] cordis.europa.eu/fp7/ict/e-infrastructure/docs/digoiduna.pdf

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[30] 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 which 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 URI's with support for content negotiation, and these HTTP URI's 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 organization that focuses on that goal as its primary mission, has a sustainable business model and a critical mass of member organizations that have agreed to common procedures and policies, has a governing body, and is committed to using open technologies.

Key aspects of the above are further discussed below.

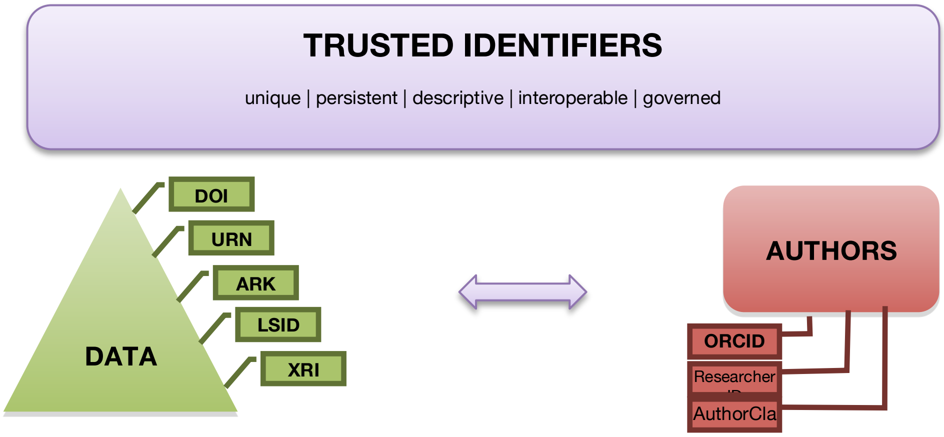


Fig. 2. The trusted identifier layer.

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

Trusted identifiers on the Web of Data

Key relevant initiatives and services in the scholarly communication space are increasingly adopting what is commonly referred to as Linked Data. The Linked Data[31] paradigm is based on a small set of core principles originally proposed by Tim Berners-Lee in 2006[32] as a pragmatic implementation of the Semantic Web ideals[33],[34]. Those principles have since 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:

- use HTTP-resolvable Universal Resource Identifiers (URIs) to identify things (both physical entities and concepts)
- when someone resolves those URIs (i.e. follows links) provide useful information in a machine-readable structured form following the standard RDF data model[35] using shared vocabularies of properties to describe those things.

Bibliographic information is increasingly available as LOD. For example, DataCite (in close collaboration with CrossRef) has for about two years returned DOI metadata as RDF in response to queries to their API, and ORCID recently implemented experimental Linked Data support in its public API.

The main hurdle to widespread use and utility of Linked Data in the scholarly domain is lack of shared ontologies of terms for describing things or, more commonly, insufficient use of existing ones. 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.

In addition to DataCite, CrossRef, and ORCID as already mentioned, other key initiatives relevant to ODIN are using or adopting LOD approaches:

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

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

[33] Berners-Lee, T., Hendler, J. & Lassila, O. The Semantic Web - A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. Scientific American 284, 34–43 (2001).

[34] Shadbolt, N., Lee, T. B. & Hall, W. The Semantic Web Revisited. IEEE Intelligent Systems 21, 96–101 (2006):.http://dx.doi.org/10.1109/MIS.2006.62

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

The Common European Research Information Format (CERIF)[36] 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 organization 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[37] and the CERIF Semantic Vocabulary[38].

The Consortia Advancing Standards in Research Administration Information (CASRAI)[39] 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[40].

The Semantic Publishing Project[41] 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[42], FaBiO[43] and FRAPO[44] 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.

In conclusion, there is clearly substantial momentum gathering around Linked Data based approaches in the scholarly communication domain. 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:

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

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

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

[39] http://casrai.org

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

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

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

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

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

- 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

Sustainability
Trusted identifiers need the strong support of an organization to become sustainable. This organization not only needs a sustainable business model, but also community support from a critical number of member organizations, and a governing body with a common set of procedures and policies.

3.3. 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 recognizes and rewards data producers[45]. 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.

Although the workflow of interoperability between data and contributor identifiers is not the focus of this document and will be addressed in a future paper, it is important to look at some key aspects of the workflow, inasmuch as they are relevant for the conceptual model. First, the information that enables data citations is created when researchers submit datasets to data centers. It is therefore crucial that data centers 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. Researchers want to get credit for these existing data publications, and for this they 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 - are described in more detail below. Both involve the same actors: the researcher, the data center, DataCite, and ORCID. The three organizations are connected to each other in a very specific way in what we call the *data citation virtuous circle*: information flows from the data center to the DataCite Metadata Store (MDS), from the DataCite MDS to the ORCID Registry, and finally from the ORCID Registry back to the data center, 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 center to ORCID when a dataset is submitted.

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

Scenario A: Submission of datasets to the data center

When a researcher deposits a dataset with a data center, his/her ORCID identifier – as well as the identifiers of other contributors – should be linked to the dataset.

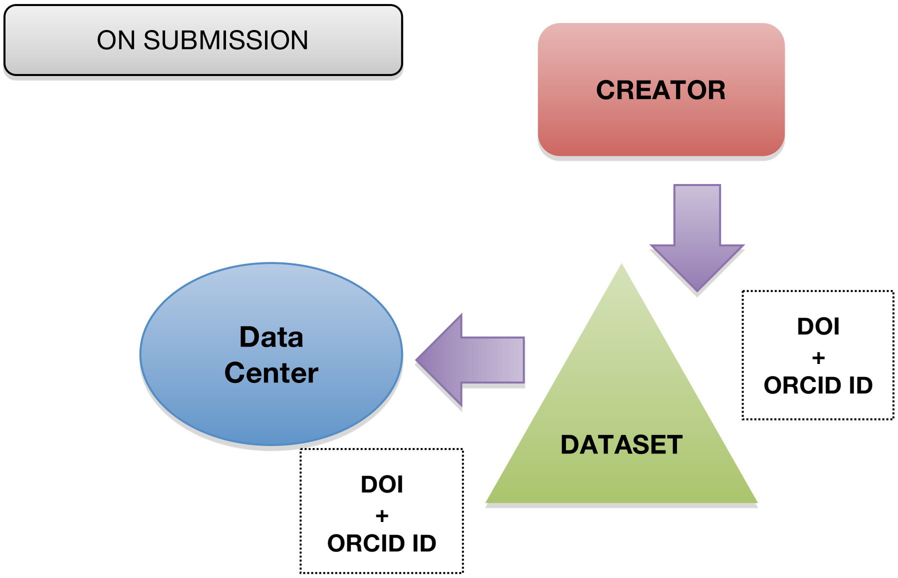


Fig. 3: Submission workflow

Other identifiers for data and/or authors can also be included in the metadata. The DataCite Metadata Schema[46] provides this functionality for data and authors. Information about the kind of contribution may also be added upon submission to the datacenter. Contributor roles currently supported by DataCite are listed in **Table 3** (see appendix).

Datasets are often part of larger collections, can exist in different versions and frequently have associated documentation and specialized software to analyze them. Another common scenario is the submission of datasets at the same time as a conventional research paper, as is standard practice for datasets submitted to the Dryad repository[47]. We want to capture all of this information in the dataset metadata, and the DataCite Metadata Schema provides properties and a vocabulary for this, for example:

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

[47] http://datadryad.org

Gugger, Paul F., Ikegami, Makihiko, and Sork, Victoria L. (2013). Data from: Influence of late Quaternary climate change on present patterns of genetic variation in valley oak, Quercus lobata Née. Dryad Digital Repository. doi:10.5061/DRYAD.G645D.

- **Has part** http://dx.doi.org/10.5061/DRYAD.G645D/1
- **Is referenced by** http://dx.doi.org/10.1111/MEC.12317

Integration of datasets and author identifiers in the data center is the biggest challenge for interoperability, because it requires a different technical implementation for every data center. To address this issue, ODIN will work on proof-of-concept technical implementations within WP3 (CERN and British Library) and WP6 (ANDS, Dryad).

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 standardized between ORCID and DataCite.

Scenario B: Claiming of already published datasets in the ORCID registry

For datasets that are already published (more than 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 for a while, until DataCite DOIs and ORCID identifiers are routinely included with datasets intended for citation.

The ORCID service was built as a self-claim system, where researchers can retrospectively claim their works. Although there are other claiming strategies, e.g., for older datasets where the researchers are no longer active, self-claiming by active researchers is the most important strategy.

Exemplifying how self-claim can work in practice, ORCID has released a service[48] 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 harmonization of the metadata formats between DataCite and ORCID. Details of this implementation work will be provided in the D4.2 end-of-project report. A similar approach is adopted by ANDS that enables linking datasets in Research Data Australia RDA)[49] to ORCID identifiers, both in the DataCite metadata and the ORCID Registry. The new service is currently in the development phase and will be available as part of the new software release (R11) of RDA (scheduled for the second half of 2013).

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

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

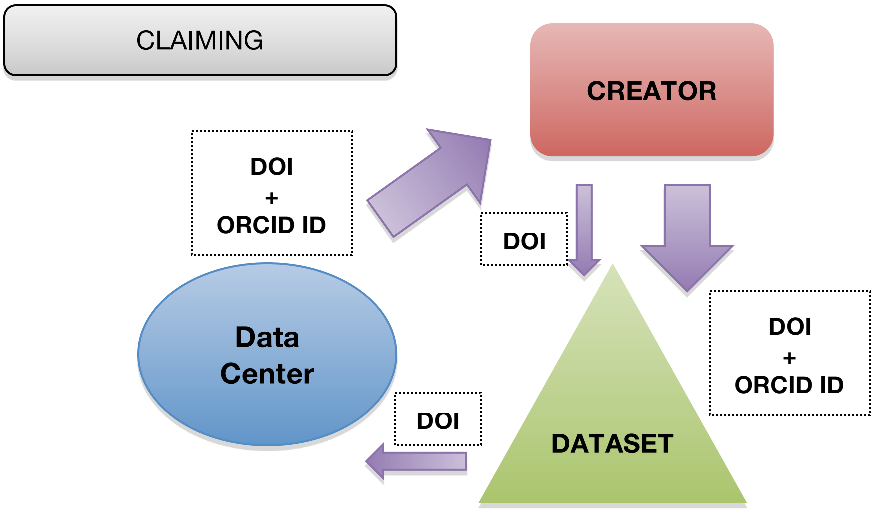


Fig. 4: Claiming workflow

Claiming can also be performed by an institution on behalf of the researcher. The workflow and tools required are the same as the self-claim case but there is an additional step 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, and all profile data marked by the contributor as publicly available are free to be reused under a CC0 waiver[50] (thus explicitly placing the data in the public domain, as per the ORCID principles[51]). The metadata also include provenance information indicating whether the claims are self-claims made by researchers, and/or claims verified by the data center.

The link between author and data also needs to be distributed back to the datacenter that created the DOI for the dataset. The data center will validate claims made by

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

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

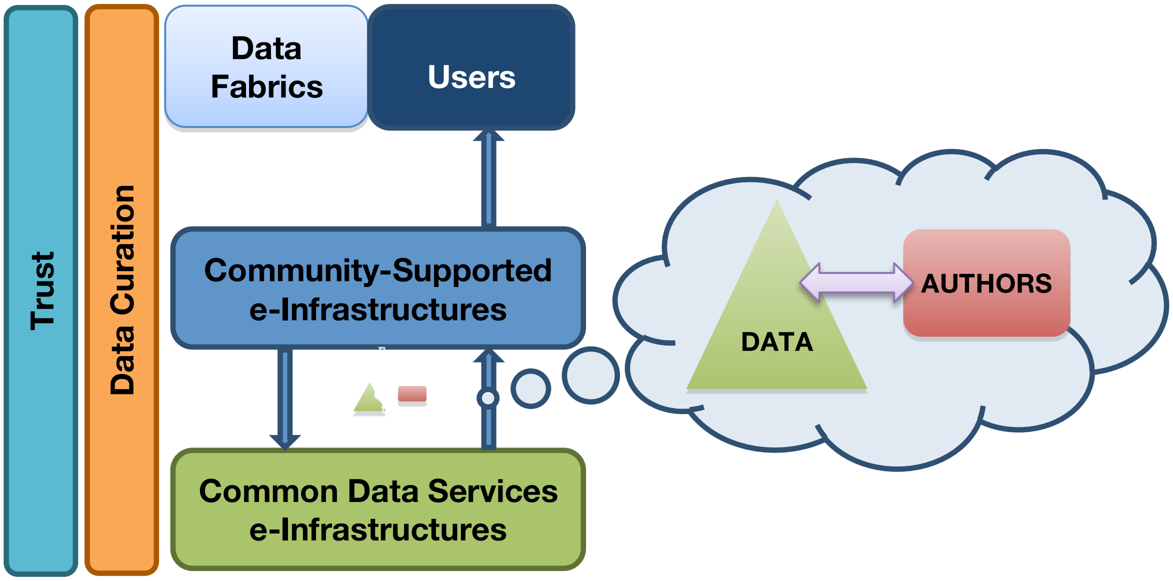
researchers and subsequently send the updated metadata to DataCite – this is the only way the DataCite Metadata Store can be updated.

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 center, adding an additional level of trust.

The technical infrastructure for the data center required to pull in claims from the ORCID registry is the same infrastructure required to pull in ORCID author identifiers (see 4.1). The ODIN partners ANDS and Dryad are working on technical proofs of concept.

3.4. 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 this interoperability layer to navigate across data and contributors. We want to shift from a paper-centric view to an interoperable layer that has expanded linkages between researchers and all of their research contributions, thereby enabling interoperability with many different services. 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. It is important that the persistent identifier layer is an independent layer not tied to related, but separate, e-Infrastructure (such as authentication or infrastructure for open access). This will decrease the dependencies on other infrastructure, 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.



Trust
Data Curation
Data Fabrics
Users
Community-Supported e-Infrastructures
Common Data Services e-Infrastructures
DATA
AUTHORS

Fig. 5: The ODIN proposal within the framework suggested by the High-Level Group on Scientific Data (The Collaborative Data Infrastructure – a framework for the future[52]).

Data and publications

One important example of interoperability beyond data and people is publications. Datasets are frequently associated with publications, and we want to navigate from datasets to publications using DOIs and other trusted identifiers, and links made between these research outputs. Although a lot of these connections between data and publications already exist, an interoperability layer of persistent identifiers will vastly facilitate making these connections, and the connections from publications to authors.

We are thus also facilitating further connections, as publications are often cited because of the data they contain, and we can link those citations back to the dataset.

Impact assessment

Scientific impact today is often demonstrated by citations, and we have scientific infrastructure to track citations to scholarly articles. The persistent identifier layer will extend this infrastructure to citations for data, and will make it much easier—and more accurate-- to track the impact of individual researchers and all their research contributions.

The interoperability layer built by DataCite and ORCID enables community-supported e-Infrastructures to share discipline-neutral, interoperable, open and persistent identifiers

[52] Riding the wave - How Europe can gain from the rising tide of scientific data - Final report of the High Level Expert Group on Scientific Data: http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf

for data and contributors. This presence of this layer will be critical for realizing the HLEG vision of riding the data wave by ensuring trust between data generators and user communities.

4. CONCLUSIONS AND NEXT STEPS

Both the unique identification of research outputs and contributors remain 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. We have created a persistent identifier layer to facilitate the interoperation of the different communities and services involved.

The end-of-project deliverable in WP2 (D4.2 Workflow for interoperability) will define more concretely how interoperable open persistent data and contributor identifiers can actually be achieved, so that users are able to navigate from a data set in a system through DataCite and ORCID to a contributor in another system. ODIN's proof-of-concept work will pave the way for implementation of a concrete workflow.

To facilitate this process, a codefest has been organized in October, in the context of the planned first year project event[53]. Among other things, the codefest will be used to gather examples that take advantage of this persistent identifier layer. The most promising proposals will be further developed as proofs of concept by the ODIN partners during the second year of the project. ODIN is working with the developer community to attract developer interest that focuses on value-added services built on top of the persistent identifier layer. These value-added services will add to its sustainability.

[53] http://odin-project.eu/events/1st-year-big-bang-and-codesprint/

APPENDIX A

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

Table 3: DataCite resource and contributor types.

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

APPENDIX B – EXAMPLES OF INTEROPERABILITY CHALLENGES

ODIN has highlighted three main threats or “items of unfinished business” emanating from lack of recognition of the need for robust ways of identifying contributors and their data in e-Science:

- Inability to follow interconnections between datasets and contributors as a method of data discovery.
- Inability to share and connect identifiers of contributors and authors between different user communities.
- Inability to uniquely identify datasets attributed to a particular contributor and contributors to a particular dataset.

These problem areas are illustrated by the following scenarios:

- 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 article, even when these data are made available by the authors, the publishers, or stored in a data repository. Once located, this data can be accessed through a community support service, and stored in a common data service, but, without a clear, open and transparent means of discovery it will remain invisible. As the published literature continues to be the primary source of inspiration for new research, this situation inhibits re-use, even within communities, and virtually prevents cross-disciplinary research. Discoverability and accessibility of underlying datasets and their creators/owners is essential for the creation of new collaborations in e-Science as well as for verification of research and the prevention and detection of scientific fraud.
- It is currently impossible to ensure that a researcher will gain scientific credit for collecting, curating and publishing datasets. The system of measuring scientific excellence is nowadays mostly built on citation measurements, focused on peer-reviewed publications. 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 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 idea of Open Data is crucial, there are often legal, ethical or commercial constraints in other fields such as privacy concerns about identifying data in medical or social science records. At present, with the high overhead of manual curation required to protect the interests of the rights holder and to check for compliance with existing regulations; the pragmatic answer is to simply deny access to problematic datasets. Without robust mechanisms for ensuring that terms and conditions of use and re-use are propagated within large-scale, automated, harvesting and data-mining tools, entire fields of scientific knowledge are rendered unusable.
- It is currently impossible to reliably connect an individual researcher, institution, region, funding agency or country uniquely to their journal articles, 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. Keeping identifier information up-to-date manually would be prohibitively expensive, while automated methods are currently unreliable. This will potentially result in significant mistrust in the e-Infrastructure.
- It is currently very difficult for institutions, funding agencies and policy makers to evaluate research projects and individual researchers. Funding organizations cannot reliably determine which research outputs they have funded, while

universities and other institutions have no reliable mechanism for monitoring the re-use of their research output. When it comes to evaluation of the full impact of the underlying data, the situation is even less satisfactory as it is impossible to track the re-use. This is a powerful disincentive for all actors to freely open exchange and re-use datasets, as that will do nothing to enhance the recognition of the value of the research.

- There are several different distributed authentication systems in use in different communities. While some of them 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 identify when two researchers co-own the same dataset, or one researcher owns two datasets. This limits the ability to realize an ecosystem in which multiple researchers asynchronously collaborate across multiple repositories.

Organization	Kind	Characteristics	Disciplines	Countries	Year started
Open Library Society	Nonprofit	Integrates with databases for institutions (ARIW) and publications (3lib.org). Started as 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 Center (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	United Kingdom	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
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 centers 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

Table 2. Contributor identifier systems overview.