DATA SCIENCE JOURNAL

Dappert, A et al 2017 Connecting the Persistent Identifier Ecosystem: Building the Technical and Human Infrastructure for Open Research. *Data Science Journal,* 16: 28, pp. 1–16, DOI: https://doi.org/10.5334/dsj-2017-028

RESEARCH PAPER

Connecting the Persistent Identifier Ecosystem: Building the Technical and Human Infrastructure for Open Research

Angela Dappert, Adam Farquhar, Rachael Kotarski and Kirstie Hewlett

The British Library, UK

Corresponding author: Dr Angela Dappert (Angela.Dappert@bl.uk)

The persistent identifier (PID) landscape extends to cover objects, individuals and organisations engaged in the process of research. Established services such as DataCite, Crossref, ORCID and ISNI are providing a foundation for a trusted ecosystem and a new generation of services. Scalable identifier systems will support researchers and capture research activity in a holistic way, across the entire lifecycle. Challenges remain – siloed services are not interoperable; important types of objects are not adequately covered, many processes remain manual, and adoption, while strong, is not consistent across disciplines.

This article draws on the work of the EU-funded THOR project to take stock of the current state of interoperability across the PID landscape and to discuss the next steps towards an integrated research record. Examples illustrate how this interconnectivity is facilitated technically, as well as social and human challenges in fostering adoption. User stories highlight how this network of persistent identifier services is facilitating good practice in open research and where its limitations lie.

Keywords: persistent identifiers; research infrastructure; interoperable services; data citation; scholarly communication

Introduction

Since the mid-1990s, persistent identifiers (PIDs) have become a key infrastructure for research. As 'an association between a character string and a resource' (DataCite Metadata Working Group, 2016), they provide significant short-term and long-term advantages. They improve the ease of locating resources; are actionable on the Web; enable metadata update and corrections without losing the resource's identity; can integrate legacy naming systems; promote linking and interoperability between services; and reduce confusion among versions of a resource. Widespread uptake of PID e-Infrastructures can accelerate the adoption of Open Science by building trust through seamless discovery of scientific artefacts; clear attribution to contributors; traceable provenance; unambiguous citation in scholarly discourse; supporting reproducibility; and enabling improved metadata quality through linking connected metadata sources. PID services, such as DataCite,[1] have a built-in governance commitment. Davidson (2006) states, 'the application of [persistent] identifiers may indicate a level of commitment on the part of the creating organisation. This can have a positive impact on the levels of trust towards that institution. Identifiers may help to provide provenance information which can positively impact the authenticity of a resource over time'. Alternative identification systems that do not support all of these characteristics have significant shortcomings. For example, URLs support access to a resource until a web site is restructured and the links break, because they identify the location rather than the resource that was once stored there.

This paper examines some of the ways in which PIDs support services beyond simple identification and enhance support for all aspects of the research lifecycle. We outline current limitations and explore

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

approaches to overcome them. Many of the technical approaches that underpin PID services are mature. As a result, much of the remaining work involves implementing higher-level functions, establishing dedicated governance structures for PID domains, and fostering societal adoption.

The benefits of achieving this vision are exemplified by a set of user stories. Some of these user stories are being examined by the Horizon2020 project Technical and Human infrastructure for Open Research (THOR). The goal of THOR is to build the technical services to support e-science and cultivate the human infrastructure required by Open Research, in particular a PID infrastructure that supports a range of intellectual endeavours beyond the borders of science. The aim is to make PIDs pervasive across platforms and stakeholders at a local, disciplinary, national and global level. Their ubiquity, interoperability and integration in diverse systems and workflows will secure a harmonised e-Infrastructure connecting artefacts and contributors. This is a crucial innovation, bridging the incomplete and fragmented PID landscape, as identified by previous work within the ORCID[2] and DataCite Interoperability Network (ODIN)[3] project.

Persistent identifiers are important

Persistent identifiers are an essential tool for resource management, but their use also benefits society. They enable global, interlinked Open Science, and ensure that the benefits of investment in research can be distributed and harvested over the long-term.

Impact

US Vice President Joe Biden (2016) has raised public awareness of the importance of data sharing as part of the US Government's 'moon-shot' for cancer initiative.[4] Cancer Research UK (2016) states: 'We must synthesise data from across disciplines to generate innovative insights into the origins, prevention and treatment of cancer. Data sharing planning is now an established part of our policies and procedures in applying for funding, and our Funding Managers and Committee members are always on the lookout for opportunities to maximise the value of our research outputs'. Data sharing is happening, but in order to meet grand challenges such as the cancer moon-shot, the sharing process must be reliable and trustworthy. Reliability and trust depend on unambiguous, accurate, and persistent identification of scientific records.

To realise the societal benefits of data sharing, it is also essential to personally empower and motivate data creators to make their data available to others. Persistent identifiers are powerful tools here too, providing a way for researchers to receive credit for their data and assert their intellectual property. This enables researchers to be less conservative and release data faster than would have been done in a traditional publication cycle. When that happens, follow-on research can be performed in more laboratories much sooner, leading to swift solutions to critical problems. This was exemplified during the 2011 E. Coli health crisis in Europe, where the rapid release of sequence data enabled swift tracking and containment of the source of the outbreak, saving lives (and businesses) in the process (Edmunds *et al.* 2012; Edmunds 2011).

User stories

It is useful to frame the benefits of persistent identifiers around those who rely on their functionality on a daily basis. Such user stories help to demonstrate current benefits and areas where services could be developed or improved.

In the following, we outline key user stories supported by persistent identifiers. Some of them are practically supported by existing services; some are technically possible but not widely implemented or supported by a managed shared infrastructure; and some need additional technical and societal solutions to be realised.

Researchers: taking credit, sharing freely

Alice wants to get appropriate credit for sharing her research data. When she submits her climate data to a trusted data centre, she includes her ORCID iD. The data centre issues a DataCite DOI for her data and ensures her ORCID iD remains associated with it. Now others can reuse her data, she gets credit, and she is still able to go on to publish papers that link back to her data (**Figure 1**).

[2] https://orcid.org/.
[3] https://odin-project.eu.
[4] https://www.whitehouse.gov/CancerMoonshot.

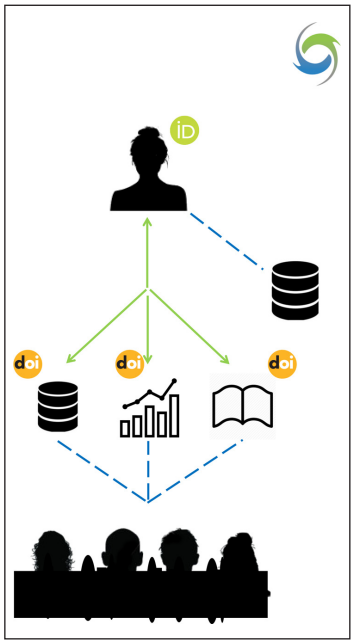

Figure 1: Researchers: taking credit, sharing freely.[5]

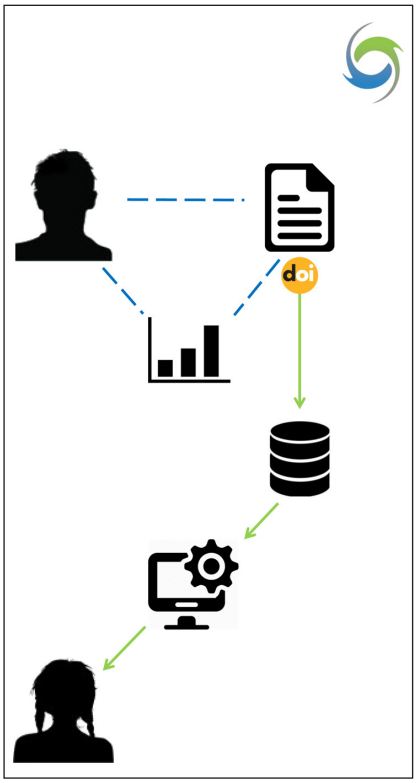

Figure 2: Researchers: discovering and reusing.

Researchers: discovering and reusing

Yannis checks out the latest publication from the Compact Muon Solenoid experiment in Physics Letters B. He wants to compare the results to a model that he has been developing. He follows a DOI link for the data behind the plots to the repository. In the repository, he can follow links to software models that use the data. One model is new to him. He follows a link to the author's profile to learn more (**Figure 2**).

[5] Image components in Figures 1, 2, 3, 4 used under license from Shutterstock.com.

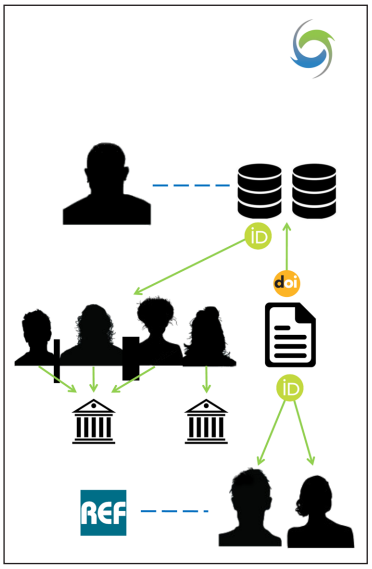

Figure 3: Data Centre Managers.

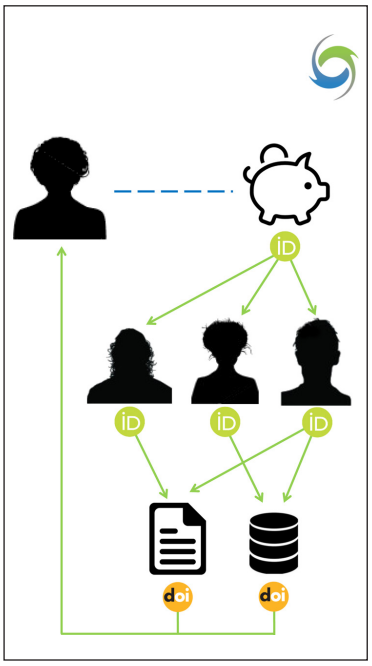

Figure 4: Funders.

Data centre managers: demonstrating value

Michele is a Data Centre Manager preparing for a bi-annual review. Success depends on demonstrating citation and re-use. As the data centre assigns DataCite DOIs to their content, they receive a notification from DataCite services whenever their data is cited. Datasets, articles, contributors, and institutions are all interlinked (**Figure 3**).

Funders: measuring impact, managing portfolios

Magda manages a portfolio of research grants. She needs to demonstrate the spread of the ideas and outputs from her portfolio. Her grant holders use ORCID iDs in her agency's systems, with publishers, and with data centres. Published outputs include an Open Funder Registry[6] ID. She can easily gather information linking her agency, grant holders, and their research outputs (**Figure 4**).

[6] http://www.Crossref.org/fundingdata/registry.html.

Publishers: including data in the publishing process

István works for a small scientific publishing company. He wants to ensure their journals publish reliable research through a transparent process. PIDs are assigned to open peer reviews and unambiguously associated with referees using ORCID iDs. Author ORCID iDs are attached to articles, which themselves have PIDs as do their data and software citations. These links make it possible to show the full provenance of each article.

Research Information Systems and Libraries: providing (long-term) access to data

Sal runs an institutional repository providing local storage and archiving for research outputs. When datasets and software held in the repository change, or related versions are deposited, she must track how each version relates to others. Each version of an item is identified by a new PID and linked through meaningful relationship types. The metadata for each version also contains the PIDs for related contributors from data creator to researcher, through to the curators and publisher, maintaining a provenance trail for each item and all its versions.

Functional requirements tied to PID architectures

The features of truly trusted and reliable identifiers that generate benefits across society go beyond the association of a resource with a character string. Following the ODIN Consortium (2013), a trusted PID must:

- Be a name, not an address
- Be globally unique
- Be persistent, designed to last beyond the lifetime of any system or (most) organisations
- Be globally resolvable as a URI with support for the full range of HTTP including content negotiation
- Be managed through a sustained committed organisation and governance process
- Come with metadata that describes its most relevant properties, including a minimum set of common metadata elements
- Be interlinkable
- Be interoperable with other identifiers through metadata elements that describe their relationships
- Be indexed and searchable by its metadata elements along with all other trusted identifiers.

Meeting all of these requirements can only be facilitated by a layered architecture (see **Figure 5**). The fundamental layer of persistent identification is currently met with identifiers such as Archival Resource

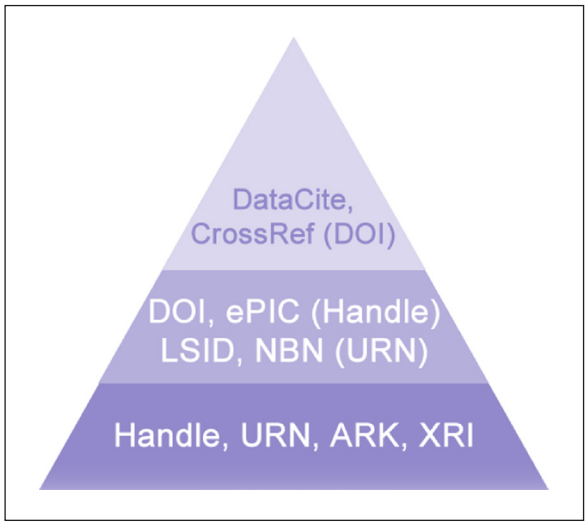

Figure 5: A representation of the layers of PID infrastructure. The basic identifier functions are supplemented with additional governance and services to support persistence.[7]

[7] URN: Uniform Resource Number; ARK: Archival Resource Key; XRI: Extensible Resource Identifier; ePIC: Persistent Identifiers for eResearch; LSID: Life Science Identifiers; NBN: National Bibliography Number.

Keys (ARKs, Kunze, 2003) and Handles,[8] which have reached a high degree of technical maturity. But these basic technical solutions do not display all the features of trusted PIDs. For example ARKs and Handles can be persistent and globally unique, they can be resolved as a URI, but they do not inherently have a committed organisation and governance process to ensure persistence; metadata is not necessarily associated with them. They can, however, be adapted and enriched with specialised services that provide added governance and functionality. This pyramid of PID solutions sees broad and fundamental PID infrastructure adapted to provide further layers of service and tailoring to specific communities.

For example, Digital Object Identifiers (DOIs) are built on top of Handle technology. The thin layer of Handle infrastructure allows the association between an identifier and a location. Additional organisations including DONA[9] and CNRI[10] provide assignment, management and maintenance services for the Global Handle Registry. The International DOI Foundation[11] (IDF) manages the DOI infrastructure and provides stronger governance to raise trust in persistence and reliability. DOIs are provided to resource owners by IDF members including DataCite and Crossref.[12] The IDF members often develop further functionality to meet the needs of a distinct stakeholder group such as data centres or publishers. Services include registration, metadata management, fragment identification, content negotiation, search and discovery, and governance. Services provided by IDF members are professionally managed and backed by a commitment to long-term persistence.

Additional requirements arise from the type of entity being identified and from the communities that they are designed to support. 'Data', as a distinct entity type, has different requirements, metadata and structural models when compared with 'publications'. Narrowing this further, subject specific data, for example biomedical data, has additional specific functions and metadata such as protocol metadata. As different communities use a PID infrastructure, they add requirements for their specific domains. The by-product of this incremental tailoring is often a narrower potential for reuse of those identifier services across other communities.

Our user stories are fundamentally enabled not just by the assignment of identifiers but the formation of links and interoperable relationships between them. This is not all currently supported to the full degree possible across all PID implementations, which leads to the proposed next steps discussed below. A deeper understanding of community practices, requirements and abilities is also needed to further expand these functionalities.

Persistent Identifier Services

PIDs for everything

The research landscape contains many types of entities that need to be identified, described, and interlinked. At a high level, however, we can group them into five categories:

1. Agents: individuals (such as researchers or curators); organisations (such as funders, research institutions, data centres, publishers and archival institutions); and other legal entities (such as consortia that are created to collaborate on research)
2. Resources: publications, data, and other research artefacts, such as lab notebooks, software, equipment, or physical specimens
3. Rights statements: grant agreements, licenses, patents
4. Events: processes that are relevant to the provenance of resources, such as creation, curation, access, claiming, updates, citation, review
5. Derived entities: such as projects, that can be seen as an aggregation of the legal entity involved, the organisation that funds it, the outputs produced and the rights statements that govern it.

Effective PID services are in place for important subsets of the first three of these categories. Current services, however, tend to focus on a single subset. For example, ORCID focuses on researchers, Crossref on articles, and DataCite on data. This makes sense from governance, business, and technical perspectives. For example, ORCID provides services to individual researchers, Crossref to publishers, and DataCite to data centres. But

[8] https://www.handle.net/index.html.
[9] https://dona.net/.
[10] http://www.cnri.reston.va.us/.
[11] https://doi.org.
[12] https://Crossref.org.

we also observe that each of these service providers also has some information about entities outside of its primary domain. For example, ORCID has links to and some information about articles and datasets associated with a researcher; Crossref on authors, funders, and related datasets; DataCite on contributors, funders, and related articles.

As a result, the data models underlying these services overlap. Further overlaps are developing as the scope of services evolve to meet expanding community needs. This leads to a need for technical and conceptual alignment.

Additionally, the layered PID architecture leads to multiple PID services that provide alternate services for an entity type, but which may address different requirements. For example, ISNI[13] and ORCID identifiers both identify individuals but meet the needs of different communities (MacEwan *et al.* 2012; Haak *et al.* 2012). The first focuses on published, possibly historic, authors of any type of material; the second focuses on current scholarly contributors. ISNIs are primarily managed by organisations, such as libraries, using what is referred to as an authority control approach. In contrast, ORCID iDs are created and owned by individuals themselves – a process that is not suited to those no longer living. Where there is overlap in the domains covered, one can achieve interoperability by declaring equality when two PIDs identify the same entity. For example, the THOR project[14] is developing software to exchange information between ISNI and ORCID when identity equality is discovered.

Agent identifiers

Agents who are individuals can be identified through schemes such as the ISNI and ORCID identifiers mentioned above. Additionally, there are identification schemes in specific domains that do not satisfy the functional requirements of PIDs. For example, the SNAP: DRGN project (Lawrence *et al.* 2015) 'is building a virtual authority list for ancient historical, mythical, and fictional agents'.

Organisations are very difficult to identify due to hierarchical internal structures, categorisation of organisations along different dimensions (e.g. taxation, legal governance, funding) and dynamic restructuring over time. Many efforts exist to identify organisations, none of which provide a satisfactory all-round solution (Bilder *et al.* 2016). A recent discussion of the issues involved has been released by the Organisation Identifier Project and is aimed at launching and sustaining a non-profit organisation identifier registry to support researcher affiliations (Cruse *et al.* 2016). Both JISC[15] and CASRAI[16] are discussing ISNI organisation identifiers that go beyond research and are sustained by national libraries (Ferguson *et al.* 2015).

Funding bodies are a special case of organisations with particular importance to research. ORCID, DataCite and Crossref enable funding awards to be linked with the contributor, data and publication entities. This data is available to funding bodies to simplify reporting processes, reduce duplicate effort, and improve understanding of the impact of research investments. As of 2016, both DataCite and Crossref provide support for funding information. Metadata that leverages the Open Funder Registry ID[17] increases interoperability and aggregation across these two service providers.

Dynamic entities, such as a consortium that executes a research project, need to be captured, but are not supported by existing PID services. Such entities are transient and may have changing membership. Providing persistent identification for them is challenging and requires effective handling of change in identity over time.

Resource identifiers

The main forms of resources currently managed with the help of PIDs are publications and data sets. Crossref is a leading PID service for publications such as journals, proceedings, books and preprints. DataCite is a leading PID service for datasets in a wide sense that may even include physical specimens.

But there are many other resources that are essential to record or access in order to maintain the global research record, such as lab notebooks, software, or physical specimens. They have varying degrees of support through PID services and existing resource PID services are increasingly adapted to support new resource types.

[13] International Standard Name Identifier (ISNI) http://www.isni.org/.
[14] https://project-thor.eu.
[15] https://www.jisc.ac.uk/.
[16] http://casrai.org/.
[17] https://github.com/Crossref/open-funder-registry.

The importance of accurate citation of data for reproducibility is matched by the importance of software citation. One piece of faulty software can impact a decade of research (for example, Eklund *et al.* 2016). Persistently identifying software and code for reproducibility raises issues that are familiar to data identification, such as versioning and granularity, but can be more complex. Not only the software, but possibly the whole creation, rendering and execution stacks under it may have to be identified in order to be able to assess the validity of data that has been created using them. They too may enter into the peer review process, especially in the software community.[18]

Additionally to employing different PID solutions for different entity types, resource producers sometimes use multiple PID types for resources depending on the degree of long-term commitment to various resources. For example, it makes sense to assign light-weight Handles to original raw data and assign DataCite DOIs to the derived, analysed data that is intended to be kept for the long-term.

Rights statements

Documents such as grant agreements, licenses and patents can be considered resources that need to be identified and managed like any other. However, they additionally serve the special role of governing the rights of agents with respect to specific resources. Because of this they are relevant in special legal, statutory and license use cases that do not apply to other resources. Because of this there is a need to manage them differently. It is especially important to persistently link rights statements with research outputs so that the access and usage conditions can be understood in an automated way. This is necessary when knowledge is increasingly accessed and processed through machine-machine transactions with minimal human intervention. Linked rights statements will also help automated analysis of compliance, for instance compliance with funding mandates for openly licenced publications and datasets. A good example for PID use for legislation can be found in EUR-Lex[19] which uses URNs as PIDs (Spinosa 2010). This European Legislation Identifier (ELI) makes it easier to access and exchange information on EU legislation.[20] Another example can be found in http://www.legislation.gov.uk/, an online resource for UK legislation made available by the National Archives of the UK, which uses three types of URI for legislation[21] namely, identifier URIs, document URIs and representation URIs (Sheridan 2010).

Event identifiers

It is increasingly understood that it is also necessary to persistently identify events that play a crucial role in the provenance of research resources and their derivative relationships. These events provide evidence of authenticity of resources, especially when resources are versioned and derived from each other.

Identifiers for derived entities

Some entities are aggregations of other entities. An example is a project that can, for example, be described by the organisation that funds it, the agents contributing to it, the outputs produced and the rights statements that govern it. There is little explicit PID support for these entities and current approaches are ad hoc.

Transcending silos

The variety of persistent identifier systems reflects differences in the types of entities being identified, community requirements, and historical factors such as when and where an approach was developed and deployed. There is not one persistent identifier system to rule them all. The emerging PID ecosystem embraces this diversity, while introducing ways to support interdisciplinary Open Science and interoperable services.

Initiatives, such as the ODIN and THOR projects are developing standards and processes to facilitate linkage and interoperation. For example, the THOR project has rationalised and aligned both metadata and vocabulary between DataCite and ORCID. Improvements have been made to DataCite and ORCID metadata and systems to support a wider range of PID types. For example, DataCite has introduced support for Funder IDs in its metadata schema (DataCite Metadata Working Group, 2016). It is now also mandatory to specify the kind of resource being identified in order to improve interoperability with

[18] http://www.artifact-eval.org/.
[19] http://eur-lex.europa.eu/homepage.html.
[20] http://eur-lex.europa.eu/eli-register/about.html.
[21] http://www.legislation.gov.uk/developer/uris.

ORCID and to improve the discoverability of research objects registered in the DataCite Metadata Store. Geospatial locations are now both human and machine readable and interoperable with standards, such as the European INSPIRE[22] regulation and Dublin Core.[23] ORCID has changed the way it manages resource types so that they can now be added flexibly. In addition, new ways of exchanging metadata have been implemented among DataCite, ORCID, Crossref, and data centres in several domains.

Figure 6 shows the current focus of linking research contributors (in particular authors), their publications, and the data that underlies the results contained in the publication. None of the three relationships in the triangle can be implemented satisfactorily without the use of PIDs. For example, links to authors, in the form of citations from their publications or data, break when author names are not unique or change over time, for example, through marriage or transliteration. Links to data (either from citations in publications to data, or from services that aggregate author research outputs) often use poor citation formats and link to broken URL locations. Links from data to publications often don't exist at all, as this information is not updated once the publication is finalised.

PIDs can address these problems. Researcher PIDs address the problems of non-unique and changing author names. One can link resources to contributors using their persistent person identifier and associate possibly changing metadata with the PID without breaking the link. PID services can include functionality, such as automatically updating relationship information between different registries (such as between Crossref and DataCite), and the ability of claiming publications or data to one's ORCID record for instance with DataCite or Crossref or publication aggregation services). This can be done by linking the relevant PIDs through defined relationships. Additionally, it can be augmented by exchanging associated metadata between the registries. These services ensure stable citation and linking mechanisms.

Beyond the links between authors, articles and data, further relationships need to be recorded in a persistent way in order to support the use cases outlined above. These may include grants and organisations such as employers and funding agencies. For global solutions, PIDs for all entities should be fully interoperable and deliver open core services for identifier assignment, look-up, resolution and mining, across geographical, temporal and organisational barriers.

While activities such as linking data sets to each other or to publications are becoming an established practice, there remain significant scalability challenges. The exchange of links and their associated metadata currently depends on numerous bilateral agreements. It is difficult for organisations to contribute links into another's infrastructure and for links to be exchanged between data centres and publishers. It is unrealistic and even undesirable to strive for one global, central infrastructure to store and propagate all links. But open distributed linking can be achieved through semantic web technology in the form

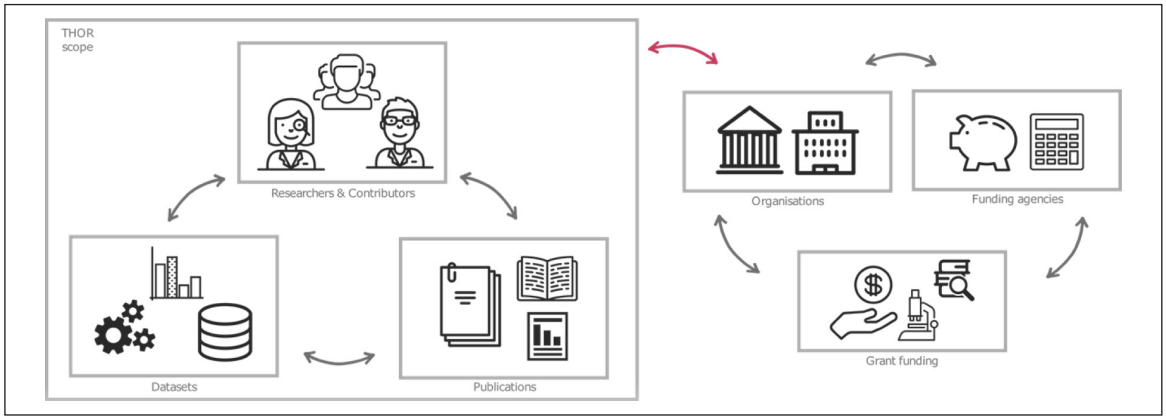

Figure 6: Transcending silos. The set of interactions on the left shows the scope of the THOR project. But the links made possible by PIDs are wider and allow for a more comprehensive view of research and its provenance.[24]

[22] http://inspire-geoportal.ec.europa.eu/.
[23] http://wiki.dublincore.org.
[24] Image by Laura Rueda with image components from flaticon.com and freepik.com.

of Linked Data.[25] For example, following the Resource Description Framework (RDF)[26] data model, links between entities are expressed as triples. Entities in the triple are identified and resolved through PIDs, such as DOIs, that can be expressed as HTTP URIs. Links between them are captured using a defined relationship type (e.g. A *is new version* of B) and provenance information. Entities and their relationships in complex data types can be structured according to schema.org[27] vocabularies. Web technology can be used to resolve the entities.

Both require an agreement across organisations on how to consistently describe and define links between research entities, and it also requires an agreement on APIs and on how to combine different sources.

Beyond the efforts of the THOR project, there are other emerging activities in this area. The RMap project, for example, addresses the need to preserve 'the many-to-many complex relationships among scholarly publications and their underlying data, thereby supporting the continual development of scholarly communication and digital publishing'.[28] RMap describes a Distributed Compound Scholarly Object, by linking PIDs of content held across a multitude of repositories to provide a single view of a research effort.

Additionally, solutions need to be implemented to preserve snapshots of linked information to ensure that it is also available for the long-term.

To support use cases such as the ones for Yannis, the researcher, and István, the publisher, bi-directional linking of datasets and publications needs to become routine practice. This not only applies to researchers' citation practice, but the workflows implemented in article submission and publication that should rely on PIDs for citation and linking, rather than fragile citations based on non-unique titles and researcher names.

Linking related resources together through their PIDs also supports better management and understanding of the provenance of research outputs. As a special case, dynamic data is data that is continually amended and added to. When a subset is extracted from a dynamic dataset, that subset needs to be persistently identified independently of the original data without necessarily having to assign a new PID for the data super-set in each case. A formal expression of the nature of the relationship of the extracted subset to the source data can be used to uniquely identify them (Rauber *et al.* 2016). There are various approaches to dealing with version control and granularity of identified objects, and linking their respective PIDs will help users to understand the relationships between those objects (Fenner *et al.* 2016).

Other types of provenance relationships are reflected in events that affect the content of a research object or that relate two objects. Going beyond citation, examples include reuse, bookmarking, online commenting, social sharing or on-line linking. Event services have recently been developed to capture these types of links. DataCite Event Data,[29] Crossref Event Data,[30] and the OpenAIRE Data/Literature Linking Service[31] collect and aggregate DOI mentions and make them available as links. These links support both Magda and Michele's use cases, forming a basis for altmetrics and impact analysis by linking research outputs to funder information and other services. CiTO, the Citation Typing Ontology, supports this sort of relationship capture by defining an ontology for describing the type of reference citation. It comprises links to other publications and also to web information resources (Shotton 2010).

Services that enable high-level cross-linking across different identifier solutions enable exchange and integration of metadata between the systems that manage those identifiers. Existing services can continue to work with internal resource or domain specific identifiers while opening up their registries for use by others, providing location-independent Linked Data support based on global resolvable PID management. One such resolving system is identifiers.org.[32] It focuses on the Life Sciences domain, supporting referencing data through URI and CURIE identifiers. Similarly, the THOR project is improving the ORCID and ISNI person identifier linking service as a basis for sharing associated works metadata between the two PID systems.

[25] The provision of resolvable identifiers (URLs) fits well with the Semantic Web vision, and the Linked Data initiative.
[26] https://www.w3.org/RDF/.
[27] http://schema.org/.
[28] http://rmap-project.info/rmap/.
[29] https://eventdata.datacite.org.
[30] https://api.eventdata.Crossref.org.
[31] http://dliservice.research-infrastructures.eu.
[32] http://identifiers.org/.

Growing scope of services

Networks of information for automatic analysis

Where linking PIDs allows us to break down silos of information, the resulting network supports the use cases that motivated our discussion. For example, data centres' need to demonstrate value and funders' need to measure the impact of their funding can be supported through better quality metrics and impact metrics (and altmetrics) based the sum of a project's outputs.

The idea of the outcomes of research being represented as a compound object, as within RMap, reduces the need for administrators to second guess what may or may not be included as an output for a particular piece of research.

A basis in the RDF data model provides an opportunity for the development of services that can automatically detect the limits of the distributed research object and disambiguate these (based on well-defined relationship types) from works that build on the original research – the works that help to demonstrate impact. Such services allow administrators such as Magda to consistently define which relationships they consider as impact, and which they do not.[33] This will require improved vocabularies to describe relationships and roles and would also be enhanced by the use of CiTO. Contributor roles in particular are not well defined. They are improving through activities such as CRediT,[34] the contributor role taxonomy, but we now have an opportunity to update this schema with one that applies to more types of output and activities that go into creating them.

Reduced administrative burden

Linking new research outputs into the network allows for services that automatically detect new contributions. Not only could this provide automatic alerts for new areas of research impact, it already reduces administrative burden by making the re-keying of information unnecessary.

For example, anyone with an ORCID iD can use DataCite's *Search and Link*[35] service to claim research objects that have DataCite DOIs and associate them with their ORCID iD. Through ORCID *auto-update*.[36] whenever a publication or a dataset receives a DOI and its metadata contains ORCID iDs, the ORCID record of the author(s) can be updated automatically. Authors receive a notification in their ORCID inbox. They can configure it to accept updates automatically in the future. This means that Alice does not need to update her ORCID record every time she publishes a paper or a dataset, and Sal does not need to manually update the university CRIS.

Persistence is built on trust

Issuing a persistent identifier for an entity is an explicit commitment to ensure the long-term availability of the item or its representation. Guarantees of long term persistence of not just the PIDs, but their related metadata are the basis for addressing Sal's use case.

Good management and governance of identifiers enables them to be persistent. On the one hand, this presents a challenge, as bad governance or organisational failure within the PID ecosystem could interrupt PID services. On the other hand, the organisations taking part in developing this ecosystem are keenly aware of this. The incorporation of long-lived institutions with clear responsibility for maintaining the research record in perpetuity, such as national libraries, along with major research institutes, adds significant strength compared to alternative approaches. Without this explicit commitment there is a reliance on location-based identification, such as URLs, which break at an alarming rate (Tyler & McNeil 2003; Rhodes 2010). But even with PIDs we see PID rot, also referred to as PID zombies (Huber & Klump 2016) when user support wanes, or when PID managing organisations disappear.

A good example of a sustainable governance migration was seen in September 2016 with the hand-over of PURL[37] governance from OCLC[38] to the Internet Archive,[39] which has a declared long-term business model.[40]

[33] For instance, they may consider A is CitedBy B to be impact for item A, but A is ReviewedBy B is not an impact.
[34] http://docs.casrai.org/CRediT.
[35] http://support.orcid.org/knowledgebase/articles/188278-link-works-to-your-orcid-record-from-another-syste.
[36] http://support.orcid.org/knowledgebase/articles/793980-what-are-auto-updates.
[37] http://www.archive.org/services/purl.
[38] http://www.oclc.org/.
[39] https://archive.org/.
[40] https://www.oclc.org/en-UK/news/releases/2016/201623dublin.html.

In the event of organisational failure, there should be mechanisms in place to recover the identifiers and the identified entities in a replacement governance structure.

While service providers strive to consolidate information across the research landscape to provide seamless and comprehensive services that avoid excessive fragmentation, there is not a reductionist drive for a single PID solution or registry. This diversity adds strength to the PID ecosystem. The resilience is enhanced when each PID service is open, exposes its information to harvesting, and registers its resources in cross-domain aggregation systems, such as GEOSS.[41] Linking and interoperability across identifier services significantly improves the value of services, as well as the potential for persistence and sustainability.

It is worth noting that the metadata associated with persistent identifiers enables linking and interoperation. To ensure the persistence of PID metadata, it should be archived and managed by organisations that have trusted, long-term metadata preservation capabilities, such as national libraries.

Where a PID identifies an information object – physical or digital – there needs to be well-defined provenance information that provides details of the history of the object. Automatic resolution to a resource needs to fail gracefully if it is no longer available. This responsibility lies with the publishers, data centres and other organisations that own the resource.

Improving adoption

Adoption of PIDs is strong. By the start of 2017, over 133 million DOIs have been created for content[42] and more than 2.5 million ORCID iDs[43] issued for researchers. As one might expect, however, this strong growth has not been uniform across disciplines and geography. By continuing to measure PID adoption, it is possible to identify areas where we can provide better services and engagement that will grow the PID landscape more equitably.

Coverage across disciplines

There is disparity in the levels of engagement with PIDs across disciplines and entity types. One data point is provided by the re3data.org[44] registry of data repositories. Of the over 1700 data repositories it lists, 99 of the 'Humanities and Social Sciences' repositories provide persistent identifiers for their data, compared to 301 of 'Life', 'Natural' and 'Engineering Sciences' repositories.[45] This difference in availability of PIDs in repositories in absolute terms may put humanities disciplines at a disadvantage for tracing the provenance of their research as well as demonstrating its impact. However, to interpret this, one would have to know what proportion of humanities and social science repositories are listed in re3data.

When taken as a proportion of subject repositories, the opposite appears true. 47% of Humanities and Social Science repositories listed provide PIDs, while just 25% of Science repositories listed use PIDs. However, availability of PIDs to disciplines may not be accurately represented in this case, as life sciences resources may hold accession numbers (such as NCBI Protein, listed at https://doi.org/10.17616/R3X039 or chEMBL listed at https://doi.org/10.17616/R3C320) rather than the PID systems prescribed by re3data ('DOI', 'ARK', 'HDL', PURL', 'URN' or 'Other'). Even though most life science accession numbers may not meet the functional requirements of PIDs as listed earlier, they can be translated into PIDs through services such as identifiers.org.

To develop a better understanding of adoption levels, as well as to track progress on closing gaps across them, will require additional data collection. Once we can effectively monitor PID usage at a disciplinary level, we will be able to develop a clearer understanding of the barriers to adoption within disciplines.

Mandates and domain specific readiness

While there is considerable growth, persistent identifier use is not yet the default. Mandates from funders (Wellcome Trust 2015; Kerridge 2015), publishers (Haak 2016) and data repositories may help to encourage wider PID use and application. This top-down approach must be complemented by ensuring

[41] http://www.earthobservations.org/geoss.php.

[42] http://www.doi.org/factsheets/DOIKeyFacts.html retrieved 2016/11/04.

[43] 2,697,767 as of 1504 GMT on 2016-11-04. From https://orcid.org/statistics.

[44] http://re3data.org.

[45] Sciences (n = 1223) and Humanities and Social Science (n = 210) repositories in re3data.org. Data provided by re3data and available from (re3data, 2016).

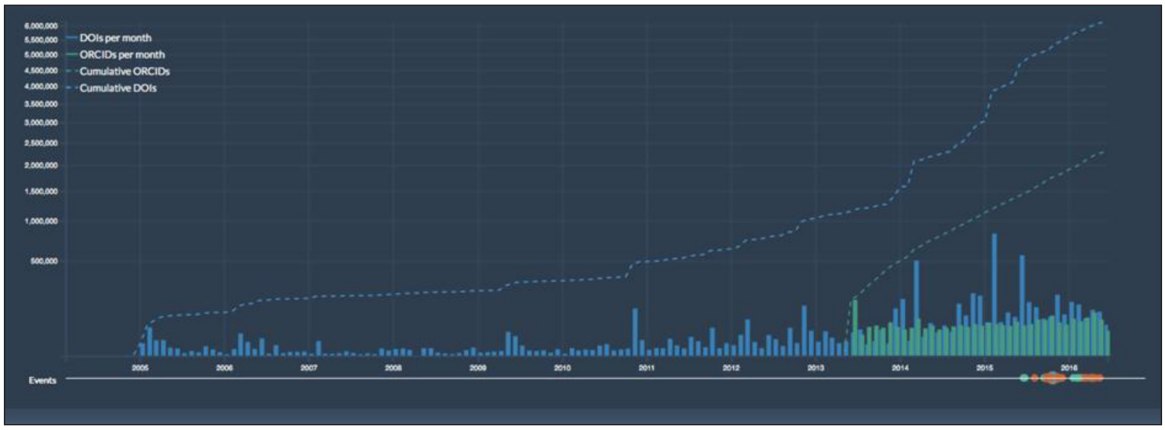

Figure 7: The THOR Dashboard showing uptake of DataCite DOIs and ORCID iDs in the years to 2016. (CrossRef, DataCite, ORCID and the THOR Project, 2016).

researchers themselves actively participate in PID infrastructures. A recent survey undertaken by ORCID (Armstrong *et al.* 2015) indicates that many scholarly researchers themselves are unaware of this trend – and even fewer are likely to have learned of ORCID as a result of mandates.

There are efforts in some disciplines to engage some parts of the research community with PID use and the benefits thereof (for example, COPDESS (2015); McNutt *et al.* (2016); Encyclopaedia of DNA Elements (ENCODE, Sloan *et al.* 2016); materials science (Austin 2016); medical sciences (Akers *et al.* 2016); crop sciences (Williams 2016); chemistry (Walter 2016); and social sciences and humanities, IMPACT-EV). Taking DOIs and ORCID iDs, we can see that once PIDs are available to the community for a given purpose, adoption naturally increases with awareness of the benefits that they bring (**Figure 7**).

Many of the advantages of PID use can only be realised once there is a critical mass of entities that have one. Sal, within the library environment, particularly needs more even uptake of PIDs across disciplines and output types to ensure what she preserves is representative of her institution.

Conclusion

Where calls by publishers, funders, scholarly societies and early-adopting researchers increase PID uptake, they may leave behind disciplines that are yet to develop the requisite community norms. To truly engage researchers in all domains on a global scale we need to think big. PIDs need to become the new normal across all disciplines and at all stages of the research lifecycle. They need to be embedded in everyday practice for everyone, from researchers, developers and librarians, to institutions, funders and data centres. That vision requires us to build successful services that utilise the power of PIDs and enhance their benefits. This may be done behind the scenes by embedding PIDs into the tools they are already using, or it may be accomplished through new services and activities that they find so beneficial they are willing to change or adapt their behaviours. PID systems and technologies must themselves adapt to the changing behaviours and practices of research in response to new technologies and processes.

Flexible and seamless integration, we believe, is the next step to ensuring adoption and long-term sustainability of PID services. Change, however, must not be to the detriment of existing users. New types of entities and functionalities may be better suited to new PID services than repurposing or diluting the focus of existing ones.

There are domains where PID use and adoption is still low, and we must investigate the factors blocking uptake and adoption and make sure that services represent the needs and activities of those areas. It may be that mandates are helpful in driving uptake, but mandates do not necessarily drive sustained use. Ultimately that will be driven by demonstrable benefits and lower barriers to use – not only to those with a vested interest in the management information that can be gained by PIDs, but to researchers in their day-to-day work.

Acknowledgements

Much of the work described in this paper was supported by the THOR project, funded by the European Commission under call H2020-EINFRA-2014-2, project number 654039. Thanks to the THOR project team for driving forward PID services, infrastructure, and use.

Competing Interests

All authors receive funding through the THOR project funded by the European Commission under call H2020-EINFRA-2014-2, project number 654039. DataCite and ORCID are THOR project partners; CrossRef and ISNI collaborate closely with the project. Adam Farquhar is co-founder and president of DataCite. Rachael Kotarski manages DataCite UK.

References

Akers, K G, Sarkozy, A, Wu, W and **Slyman, A** 2016 ORCID Author Identifiers: A Primer for Librarians. *Medical Reference Services Quarterly*, 35(2): 135–144. DOI: https://doi.org/10.1080/02763869.2016.1152139

Armstrong, D, Haak, L, Meadows, A and **Stone, A** 2015 ORCID 2015 Survey Report. DOI: https://doi.org/10.6084/m9.figshare.2008206

Austin, T 2016 Towards a digital infrastructure for engineering materials data. *Materials Discovery*. DOI: https://doi.org/10.1016/j.md.2015.12.003

Biden, J 2016 Cancer Moonshot. Report to the President from the Vice President. *Washington: The White House*. Retrieved from: https://www.whitehouse.gov/sites/default/files/docs/finalvp_exec_report_10-17-16final_3.pdf.

Bilder, G, Brown, J and **Demeranville, T** 2016 Organisation identifiers: current provider survey. DOI: https://doi.org/10.5438/4716

Cancer Research UK 2016 Why aren't we sharing? London: *Cancer Research UK*. Retrieved from: http://www.cancerresearchuk.org/funding-for-researchers/research-features/2016-08-10-why-arent-we-sharing.

COPDESS 2015 COPDESS Statement of Commitment – COPDESS. Retrieved from: http://www.copdess.org/statement-of-commitment/ 1 Nov. 2016.

CrossRef, DataCite, ORCID and the THOR Project 2016 THOR Project Dashboard Overview. The graph is generated from data retrievable from http://dashboard.project-thor.eu/api/data?type=aggregate. The graph image was accessed from http://dashboard.project-thor.eu/dashboard/ December 2016.

Cruse, P, Haak, L and **Pentz, E** 2016 Organization Identifier Project: A Way Forward. DOI: https://doi.org/10.5438/2906

DataCite Metadata Working Group 2016 DataCite Metadata Schema Documentation for the Publication and Citation of Research Data. Version 4.0. DataCite e.V. DOI: http://doi.org/10.5438/0012

Davidson, J 2006 Persistent Identifiers. DCC Briefing Papers: Introduction to Curation. *Edinburgh: Digital Curation Centre*. Handle: 1842/3368. http://www.dcc.ac.uk/resources/briefing-papers/introduction-curation/persistent-identifiers#sthash.jpZBaD5T.dpuf.

Edmunds, S 2011 Notes from an E. coli "tweenome" – lessons learned from our first data DOI. Retrieved from: http://blogs.biomedcentral.com/gigablog/2011/08/03/notes-from-an-e-coli-tweenome-lessons-learned-from-our-first-data-doi/ 4 Dec. 2016.

Edmunds, S, Pollard, T, Hole, B and **Basford, AT** 2012 Adventures in data citation: sorghum genome data exemplifies the new gold standard. *BMC Research Notes 2012*, 5: 233. DOI: https://doi.org/10.1186/1756-0500-5-223

Eklund, A, Nichols, T E and **Knutsson, H** 2016 Cluster failure: why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences*, p. 7900. DOI: https://doi.org/10.1073/pnas.1602413113

Fenner, M, Demeranville, T, Kotarski, R, Dasler, R, McEntyre, J, de Mello, G, Vision, T, Dappert, A and **Farquhar, A** 2016 THOR: Conceptual Model of Persistent Identifier Linking. DOI: https://doi.org/10.5281/zenodo.48705

Ferguson, N, Moore, R and **Schmoller, S** 2015 Review of selected organisational IDs and development of use cases for the Jisc CASRAI-UK Organisational Identifiers Working Group. Retrieved from: http://repository.jisc.ac.uk/id/eprint/5853 20 Dec. 2016.

Haak, L 2016 Publishers to Start Requiring ORCID iDs [Text]. Retrieved from: https://orcid.org/blog/2016/01/07/publishers-start-requiring-orcid-ids 5 May 2016.

Haak, L L, Fenner, M, Paglione, L, Pentz, E and **Ratner, H** 2012 ORCID: a system to uniquely identify researchers. *Learned Publishing*, 25(4): pp. 259–264. DOI: https://doi.org/10.1087/20120404

Huber, R and **Klump, J** 2016 How dead is dead in the PID Zombie Zoo? Retrieved from: https://www.rd-alliance.org/sites/default/files/attachment/20160902-RDA_EU_View_on_PID_Systems_Garching-Robert_Huber-Jens_Klump-How_dead_is_dead_in_the_PID_Zombie_zoo.pdf 23 Dec. 2016.

Kerridge, S 2015 The UK Recommends ORCID [Text]. Retrieved from: http://orcid.org/blog/2015/07/16/uk-recommends-orcid 5 May 2016.

Kunze, J 2003 Towards electronic persistence using ARK identifiers. In *Proceedings of the 3rd ECDL Workshop on Web Archives*. Retrieved from: https://wiki.umiacs.umd.edu/adapt/images/0/0a/Arkcdl.pdf 20 Dec. 2016.

Lawrence, K F and **Bodard, G** 2015 June. Prosopography is Greek for Facebook: The SNAP: DRGN Project. In *Proceedings of the ACM Web Science Conference*, p. 44. ACM. DOI: https://doi.org/10.1145/2786451.2786496

MacEwan, A, Angjeli, A and **Gatenby, J** 2012 The International Standard Name Identifier (ISNI): The evolving future of name authority control. *Cataloging & Classification Quarterly*, 51(1–3): pp.55–71. DOI: https://doi.org/10.1080/01639374.2012.730601

McNutt, M, Lehnert, K, Hanson, B, Nosek, BA, Ellison, AM and **King, JL** 2016 Liberating field science samples and data. *Science*, 351(6277): 1024–1026. DOI: https://doi.org/10.1126/science.aad7048

ODIN Consortium, Fenner, M, Thorisson, G, Ruiz, S and **Brase, J** 2013 D4.1 Conceptual model of interoperability. figshare. DOI: https://doi.org/10.6084/m9.figshare.824314.v1

Rauber, A, Asmi, A, van Uytvanck, D and **Pröll S** 2016 Identification of Reproducible Subsets for Data Citation, Sharing and Re-Use. Retrieved form: https://www.rd-alliance.org/system/files/documents/TCDL-RDA-Guidelines_160411.pdf Retreived 01 Dec. 2016.

re3data 2016 Re3data API with documentation available from http://www.re3data.org/api/doc. Data extracted 15 November 2016.

Rhodes, S 2010 Breaking down link rot: the Chesapeake project legal information archive's examination of URL stability. *Law Libr. J.*, 102: p. 581.

Sheridan, J 2010 Legislation.gov.uk. VOXPOPULII. Retrieved from: https://blog.law.cornell.edu/voxpop/2010/08/15/legislationgovuk/ 23 Dec. 2016.

Shotton, D 2010 CiTO, the Citation Typing Ontology, *Journal of Biomedical Semantics*. Springer Nature. DOI: https://doi.org/10.1186/2041-1480-1-s1-s6

Sloan, C A, Chan, E T, Davidson, J M, Malladi, V S, Strattan, J S, Hitz, B C, Cherry, J M et al. 2016 ENCODE data at the ENCODE portal. *Nucleic Acids Research*, 44(D1): D726–D732. DOI: https://doi.org/10.1093/nar/gkv1160

Spinosa, P, Francesconi, E and **Lupo, C** 2010 A uniform resource name (URN) namespace for sources of law (LEX). Retrieved from: https://tools.ietf.org/id/draft-spinosa-urn-lex-05.txt 23 Dec. 2016.

Tyler, D C and **McNeil, B** 2003 Librarians and link rot: A comparative analysis with some methodological considerations. *portal: Libraries and the Academy*, 3(4): pp. 615–632. DOI: https://doi.org/10.1353/pla.2003.0098

Walter, P 2016 2 February. I am a number! Retrieved from: http://www.rsc.org/chemistryworld/2016/02/researcher-number-orcid-news-leader-editorial.

Wellcome Trust 2015 Who are you? Recognising researchers with ORCID identifiers. Retrieved from: https://blog.wellcome.ac.uk/2015/06/30/who-are-you-recognising-researchers-with-orcid-identifiers/.

Williams, S C 2016 Practices, Policies, and Persistence: A Study of Supplementary Materials in Crop Science Journals. *Journal of Agricultural & Food Information*, 17(1): 11–22. DOI: https://doi.org/10.1080/10496505.2015.1120213

How to cite this article: Dappert, A, Farquhar, A, Kotarski, R and Hewlett, K 2017 Connecting the Persistent Identifier Ecosystem: Building the Technical and Human Infrastructure for Open Research. *Data Science Journal,* 16: 28, pp. 1–16, DOI: https://doi.org/10.5334/dsj-2017-028

Submitted: 11 January 2017 **Accepted:** 26 May 2017 **Published:** 15 June 2017

]u[*Data Science Journal* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS