Project Name	**FREYA**
Project Title	**Connected Open Identifiers for Discovery, Access and Use of Research Resources**
EC Grant Agreement No	**777523**

D3.1 Survey of Current PID Services Landscape

Deliverable type	Report
Dissemination level	Public
Due date	31 May 2018
Authors	Christine Ferguson, Jo McEntrye (EMBL-EBI) Vasily Bunakov, Simon Lambert (STFC) Stephanie van der Sandt (CERN) Rachael Kotarski, Sarah Stewart, Andrew MacEwan (BL) Martin Fenner, Patricia Cruse (DataCite) René van Horik (DANS) Tina Dohna, Ketil Koop-Jacobsen, Uwe Schindler (PANGAEA) Siobhan McCafferty (ANDS)
Abstract	A comprehensive survey of the landscape of persistent identifiers across many disciplines is presented, with assessments of maturity of different PID types and conclusions for the future.
Status	Submitted to EC 17 July 2018

The FREYA project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 777523.

FREYA project summary

The FREYA project iteratively extends a robust environment for Persistent Identifiers (PIDs) into a core component of European and global research e-infrastructures. The resulting FREYA services will cover a wide range of resources in the research and innovation landscape and enhance the links between them so that they can be exploited in many disciplines and research processes. This will provide an essential building block of the European Open Science Cloud (EOSC). Moreover, the FREYA project will establish an open, sustainable, and trusted framework for collaborative self-governance of PIDs and services built on them.

The vision of FREYA is built on three key ideas: the **PID Graph**, **PID Forum** and **PID Commons**. The PID Graph connects and integrates PID systems to create an information map of relationships across PIDs that provides a basis for new services. The PID Forum is a stakeholder community, whose members collectively oversee the development and deployment of new PID types; it will be strongly linked to the Research Data Alliance (RDA). The sustainability of the PID infrastructure resulting from FREYA beyond the lifetime of the project itself is the concern of the PID Commons, defining the roles, responsibilities and structures for good self-governance based on consensual decision-making.

The FREYA project builds on the success of the preceding THOR project and involves twelve partner organisations from across the globe, representing PID infrastructure providers and developers, users of PIDs in a wide range of research fields, and publishers.

For more information, visit www.project-freya.eu or email info@project-freya.eu.

Executive summary

Partner organisations involved in the FREYA project were briefed to provide, within the first six months of the project, an assessment of the landscape of established and emerging persistent identifiers (PIDs) used in scholarly research at the time. This is important because seeking PID maturity (consensus around what specifically needs identifying and which PID type is most relevant to assign to an entity), linking PIDs across entities; and growing overlay services and the PID infrastructure will enable increased discoverability of research outputs themselves, as well as an understanding of the return on research investments.

FREYA partners include “big data” producers and handlers as well as publishers, and represent a diverse array of research disciplines including the life sciences, physical, environmental and social sciences, arts and humanities, in Europe and more globally.

The evaluation has been distilled into a report comprised of a landscape scan and a maturity matrix evaluation of PID services. More specifically, the maturity table lists research resources (entities), the identifier types currently assigned to these categories of research resources, and a high-level assessment of the maturity of associated services. Currently only three entities (researchers, publications and data) have services that are deemed fully mature. These entities have been the focus of projects that preceded FREYA, namely ODIN (ORCID and DataCite Interoperability Network) and THOR (Technical and Human infrastructure for Open Research). There is growing interest from a number of stakeholders to further address the need for open and global identifier systems for entities such as organisations, grants, software, research facilities and conferences—PID services for these entities have been categorised as “emerging”. PIDs and interlinking services around entities such as field stations, cultural artefacts and certain types of study registrations are deemed immature. Following the table, detailed summaries of the PID landscape for each research object are provided, plus discussion of the challenges that block elevation to a higher level of maturity. The report concludes with general observations for readers about PID usage across the ecosystem.

Beyond a benchmarking document for partners at the outset of project FREYA, the report is intended to be a reference document for stakeholders across disciplines seeking a summary of the current identifiers in use and the extent to which cross-linking of entities is possible. Although FREYA partners are well-placed to draw up this initial report, feedback from readers is encouraged to amplify and refine any of the assessments made.

Contents

1 Introduction

A well-developed infrastructure for certain types of persistent identifiers (PIDs) already exists and is widely used for referencing publications, datasets and individual researchers. DOIs (digital object identifiers) are commonly used to reference research publications, ORCID iDs are an identifier type increasingly being assigned to researchers, while datasets are currently referenced by a host of different identifiers including DOIs, accession numbers, URLs, etc[1]. These identifiers are important elements of the fabric of scientific research, but there are others, and there are many ways that these elements may be linked together to add value in the evolving e-infrastructure of the European Open Science Cloud (EOSC) and beyond. For example: services exist that enable a researcher (via his/her ORCID iD) to claim their research outputs by linking directly to specific publication DOIs and dataset identifiers. Beyond assigning and linking these PIDs, an assessment of performance/usage of these outputs is key, for example measuring impact through citations and other contributions to the advancement of science and society. The goal of FREYA is to extend the scope of PIDs and the services operating over these identifiers. One aspect involves extending the range of research-relevant entities that can be identified and referenced with PIDs; the other is expanding what can be done with those PIDs.

Almost any class of entity can be individuated, whether a thing in the real world, digital object or abstract concept. These entities can also potentially be associated with persistent identifiers. However, there has to be some limitation on FREYA's ambition. That limit is determined by what is relevant and useful to European and global research, where usefulness is understood in terms of enriching context: the value-added creation and expansion of the "PID Graph" that is one of the driving visions of FREYA. The *PID Graph "*connects and integrates PID systems, creating relationships across a network of PIDs and serving as a basis for new services"[2]. Beyond the trio of publication–dataset–person mentioned above, work is already under way in some areas to define new PID types, implement them with existing or new technological bases, and use them in applications. Some of these developments are quite generic—for example, PIDs for organisations or software—while others are specific to the needs and practices of particular research fields—it has been suggested, for example, that historical and mythical personages could be associated with PIDs. FREYA must balance its own developments with channelling and leveraging what is already progressing—there is no sense in duplication of effort or needlessly inventing competing approaches.

The present deliverable does the groundwork for understanding the landscape of these developments with a view to focussing and prioritising the work of FREYA. This is envisaged in the project concept which comprises three "pillars": the aforementioned PID Graph, the PID Forum of stakeholders and the PID Commons, which is the open framework for collaborative self-governance of the PID infrastructure. FREYA's engagement with a wider range of PID entities is framed in terms of Technology Readiness Levels: "New services, and new PID types, will be introduced and moved up the scale of Technology Readiness Levels". This has implications for assessing and planning for the maturity of the technology, especially the dimension of validation and demonstration.

It is vital to distinguish new entities that may be identified with PIDs; new types of PID technology (which are not necessarily required, as for instance many different entities may be assigned a DOI or a Handle); and new services acting over them. Services may be at three levels: basic services for issuing and resolving PIDs; linking services that effectively create the PID Graph; and community services that create meaningful aggregations and add value through traversing the graph. Such value may arise from tracking provenance, or by enabling metrics for open science.

The position of this deliverable in the FREYA project is therefore to survey and assess the current state of the art for PIDs and PID services, in terms of gaps, that is, what is lacking. Gaps may arise with regards to missing functionality of existing PID services, and concerning the lack of available PID services for important

[1] For full definitions of identifier types, see Identifier abbreviations in Annex A.
[2] https://www.project-freya.eu/en/about/mission

scholarly resources. These are interpreted as maturity gaps, in as much as they reflect a need that is yet unmet by the available offerings. It follows that it is necessary to have in mind some description of potential use in order to provide context to the PID types and services being discussed. The following deliverable D3.2 is the place for use cases for specific service improvements, developments, requirements, and prioritization of work arising as a consequence.

The expectation is that by the end of this exploration of new PID types in the FREYA project, prototype services arising from this work will either be taken up in the work packages concerned with core PID services or (if domain-specific) integrating the PID Graph, or further developed outside of FREYA, or documented and sunsetted at the end of the project.

From a reader's point of view this document should therefore be seen as a reflection of the awareness of the FREYA partners on developments in new PID types; an overview with a rough maturity evaluation of these developments; and an opportunity to feed back, amplify and correct any of the assessments made[3].

[3] Please email the FREYA project via info@project-freya.eu

2 Methodology

2.1 Assessment approach

This section describes the approach taken to assess the PID landscape for this report and notes any relevant caveats.

The combined knowledge of FREYA partners with additional conceptual input taken from early discussions with the community, such as conversations at the Research Data Alliance, (RDA) and presentations at the PIDapalooza meeting in Girona 2018[4], was used to collate a view of the current status of the PID ecosystem and how it is evolving in the immediate term.

The process involved the following steps:

1. A series of group discussions to identify a list of research object types or entities relevant to research for which identifiers may be required or in use. These are summarised in Table 1. Group discussions were also held to capture what is understood by the term "maturity" and how it relates to different components of the PID landscape: see more on this below.
2. Partners familiar with the status of PID services associated with specific research entities then provided an description of the relevant portion of the PID ecosystem and an assessment of its "maturity". These descriptions reference published material where possible, but are necessarily current in that they point to more informal blogs, conference reports or summaries of working group discussions.
3. Grouping the research entities according to similarity/broad themes. With an understanding of what a specific research entity means to different disciplines, the entities were grouped according to themes of commonality among entities and to avoid "special cases" where possible. The entities discussed in this report are listed and grouped by broad theme in Table 1.
4. Sanity-checking the final document via internal review i.e. in consultation with FREYA partners who had not been directly involved in the initial drafting.

2.2 Defining "maturity"

In this report, the FREYA partners aimed to impart a relative sense of the maturity of different parts of the PID ecosystem. Agreeing a definition of maturity turned out to be challenging since it required clarifying what exactly is being measured and what relative scale should be used as a proxy for maturity vs immaturity.

What is being measured could relate to the research resource/entity: maturity could be a measure of whether an entity is yet assigned identifiers; how extensively the research entity is linked to other entities; the uptake of identifier use for an entity within a research discipline versus across research disciplines. It could also relate to the PID type per se: here maturity could be a measure of time—how long the PID type has been used to identify the research entity; alternatively, a measure of how many PIDs are assigned to an entity (e.g. the number of ORCID iDs vs number of ISNIs to identify researchers); how many research entities use a particular type of PID (e.g. DOIs are used as identifiers for several research entities yet ORCID iDs are restricted for use as people identifiers).

We agreed to focus on the research entity for the maturity matrix presented in Table 1, and for each to provide a sense of how advanced the *identifier infrastructure* is around that entity.

[4] https://www.project-freya.eu/en/news/newsitems/blog-pidapalooza

The following terms and definitions were considered for the maturity rating,

Mature: infrastructure in common research community use; regular use within a research discipline (would have to specify whether >1 research discipline employs the infrastructure).

In pilot: some demos, infrastructure not in common research community use.

Emerging: PID forum discussions or specific working groups convened, infrastructure/services being actively planned. Alternatively, infrastructure is established (perhaps because it uses an existing infrastructure such as DOI) but the PID has not gained much traction yet, or there are a diversity of approaches with little clarity (or consensus) on which to use in any given situation.

Immature: nascent PID forum discussions, or no clarity on whether or how to model the research object (entity) as a PID.

The maturity scale that was employed in Table 1:

Mature: infrastructure in common research community use i.e. regular use within a research discipline ;

Emerging (to incorporate those that have services "*In Pilot"*) : infrastructure not yet available for common research community use; services may be in pilot or being actively planned by a working group.

Immature: nascent PID forum discussions; no definite consensus.

The rationale for our focus on scoring the PID infrastructure around research entities using this rating is as follows: some of the aims of describing the PID landscape is to provide a general overview, and highlight gaps in the PID ecosystem that could be filled. Where a gap is identified for an entity the aim is to map it against the existing ecosystem and use the draft maturity matrix to determine whether there are pre-existing PID services of the type that might be employed to fill a gap or an existing working group whose relevant expertise could be tapped.

The "maturity rating" presented in Table 1 has several caveats given the specific focus on PID services—the terms do not provide specifics about the breadth of uptake of the PIDs or services across different disciplines e.g. whether a PID is employed in any or all of the following: life sciences, humanities and social sciences, physics, geosciences. The table does not offer sheer numbers of a specific PID type assigned to any given entity. These numbers would not be comparable for the same PID type used across different disciplines: the datasets in the Life Sciences are different in nature to those collected in High Energy Physics and Humanities domains; datasets in different disciplines make use of different and often field-specific PID types; numbers of datasets will vary from one discipline to another.

To address the above caveats, nuance and granularity is added to the specific sections that follow Table 1 offering more in-depth discussion for each research entity type. The concluding section of this document includes more general observations about PID types in current usage.

3 Results

3.1 Overview of PID types in usage and the maturity of respective PID infrastructure

Table 1 presents a list of entities (i.e. research object types) for which PIDs are in use (even if at a small scale), or are being/have been considered. For each entity, PID types in use and estimated maturity of the PID services/infrastructure is provided. The entities are grouped in themes, marked by specific colours. A more comprehensive description of the current state of PID ecosystem for each research entity, may be found in the sections following the table.

Table 1 Entities, PID types and their maturity

Research entity	**PID types used**[5]	**Maturity of PID Infrastructure**
Publication	DOI, Accession number, Handle, URN, Scopus EID, Web of Science UID, PMID, PMC, arXiv Identifier, BibCode, ISSN, ISBN, PURL	Mature
Citation	OCI (secondary aggregation of information)	Emerging
Conference	DOI, Accession number	Emerging
Researcher (or Scholar)	ORCID iDs, ISNI (also DAIs, VIAFs, arxivIDs, OpenIDs, ResearcherIDs, ScopusIDs)	Mature
Organization	DOI; ISNI, GRID, Ringgold IDs	Emerging
Data	DOI, Accession number, Handle, PURL, URN, ARK	Mature
Data repository		Immature
Grants	DOI, PURL	Emerging
Project	local identifier, accession number, RAiD	Emerging
Experiment	none	immature

[5] For full definitions of identifier types, see Identifier abbreviations in Annex A.

Investigation	DOI, Accession number	Emerging
Analysis	Git gist	Immature
Software	DOI, SHA-1 hash	Emerging
Computer Simulation	UUID	Emerging
Software License	none	Immature
Equipment		
Instrument, Device, Sensor, Platform, Research Facility	DOI, RRID, UID	Emerging
Archival/Storage facility	URI, DOI, UUID	Emerging
Field Station	none	Immature
Sample		
Geological or Biological Sample	Accession number, RRID, DOI, IGSN	Emerging
Cultural artefact	DOI, URN, Accession number	Emerging
Historical or mythical person	URI	Emerging
Temporal period & historical place	ARK, URI, accession number	Immature
Study registration		
Clinical trial; non-clinical registration	accession number; DOI	Immature
Data Management Plan	DOI	Immature
Workflow	URI, DOI	Immature
Protocol	DOI	Immature

3.2 PID usage around specific information entities

The sections below expand on the summary provided in the table above. Each section includes a definition of the research entity and what might it mean to different disciplines; an indication of why the research object type is included in this report (e.g. raised in discussions RDA, discussions with disciplinary researchers); a description of the maturity gaps—the issues/challenges that are blocking elevation to a higher level of maturity; an indication of whether services for this PID type are within reach of working groups including FREYA partners.

3.2.1 Publications

Research articles, books, preprints and theses are some of the many publication types referred to in this section[6]. Research articles are one of the oldest and most common scientific resources and a fundamental part of scholarly communication. As a well established expression of scholarly outcome, research articles share a long tradition of using PID systems. The market for scholarly publications is still growing and with a highly increasing number of publishers, information infrastructures and academic journals, the demand for unique and persistent identification of research objects is higher than ever. Referencing research articles is an essential part of good scientific behaviour, but URL decay ("link rot"[7]) or "content drift"[8] has made it very difficult for users to refer to digital objects in a persistent way and made many resources inaccessible to others. The implementation of PID systems solved this problem for scholarly communication and enabled a precise and unambiguous identification of resources.

There are a variety of PID types in circulation and the choice of PID system depends on provider and disciplinary habits. But as demonstrated in the THOR project, the DOI is the most established system in use for research articles[9]. Looking at the distribution of PID systems in ORCID author records, DOI is the primary PID type followed by popular database-specific identifiers, such as Scopus EID, Web of Science UID and PubMed (PM) ID for abstracts. PID types such as ISSN, ISBN, PubMed Central (PMC) identifiers (for full text articles), arXiv, BibCode and Handle identifiers are also well established, though not as widely used.

Usage, referral to or citations of published scholarly articles, can be considered as data entities per se, not only as links between published entities. OpenCitations[10] is a small initiative working to put in place a common ontology for machine readable definition, a persistent identifier scheme (Open Citation Identifier, OCI) and resolution service infrastructure - these comprise requirements for citations to be treated as "first class data entities".

3.2.2 Conferences

In certain disciplines. e.g. computer science, a significant proportion of the research is published not in journals, but in conference proceedings. In other disciplines conference presentations and posters are an essential part of the scientific discourse. While conference proceedings may be published using a persistent identifier, there is currently no way to unique associate these outputs with a conference. To address this gap, a joint Crossref/DataCite working group was started in February 2017[11], based on initial work by

[6] See publication types here : https://europepmc.org/advancesearch
[7] Klein M, Van de Sompel H, Sanderson R, Shankar H, Balakireva L, Zhou K, et al. (2014) Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot. PLoS ONE 9(12): e115253. https://doi.org/10.1371/journal.pone.0115253
[8] Jones SM, Van de Sompel H, Shankar H, Klein M, Tobin R, Grover C (2016) Scholarly Context Adrift: Three out of Four URI References Lead to Changed Content. PLoS ONE 11(12): e0167475. https://doi.org/10.1371/journal.pone.0167475
[9] Dallmeier-Tiessen, Sünje, & Dasler, Robin. (2016, September 21). Analysis and Comparison of Persistent Identifier Use and Integration across Disciplines and Sectors. Zenodo. http://doi.org/10.5281/zenodo.154592
[10] www.opencitations.net
[11] https://www.crossref.org/blog/taking-the-con-out-of-conferences/

SpringerNature. The working group subsequently produced an initial specification for conference identifiers and associated metadata, that was open for public comment until the end of May 2018[12].

The work on conference identifiers encompasses two related activities: a) describing conference series and individual conferences using appropriate metadata, and b) describing the conference outputs, with a focus on conference proceedings. The working group has identified a number of important use cases, has seen strong interest in the Crossref and DataCite communities, and is now moving into the implementation phase, finalizing the metadata, and discussing governance and business models.

3.2.3 Researchers

The term "researcher" is used inclusively here for investigators, scholars, authors, creators and researchers. Persistent Identifiers have been used to uniquely identify and disambiguate individual researchers, thereby acting as an authority control, linking creators with their works, and also solving issues with people having identical names in the scholarly record, or those that change over time to reflect a change in marital status for example. Persistent identifiers for people include standards such as ISNI (International Standard Name Identifiers) and ORCIDs (Open Researcher and Contributor IDs). There are other identifiers that relate to authority files for specific uses, such as DAIs (Digital Author Identifiers), VIAFs (Virtual International Authority Files) and arXiv author identifiers. There are also proprietary persistent identifiers such as ResearcherIDs (Clarivate Analytics, formerly Thomson Reuters) and ScopusIDs (Elsevier Scopus). "Researchers" in the context of persistent identifiers therefore refer to an authority control for individuals who have authored or created works, whether they are academic researchers, authors, artists, musicians or scientists. Persistent identifiers for individuals are therefore essentially pan-disciplinary, although the uptake of some forms of persistent identifiers may be geared towards researchers working in an academic research milieu, with mandates from funders, publishers and institutional systems such as Central Research Information Systems (CRIS) playing key roles in the uptake of these forms of identifiers.

Persistent identifiers for people may have differing roles in different disciplines. In science and technical disciplines, there has been a high uptake of ORCID iDs, with clinical medicine, technology and applied sciences and biology leading with the highest number of ORCID iDs and linked datasets[13,14]. Uptake of ORCID iDs is much lower in the arts, humanities, economics and social sciences, although the general uptake of ORCID iDs in the humanities has increased gradually since 2012, particularly in the social sciences. The difference in uptake between the sciences and the arts may be explained in part through mandates from funders and publishers, in addition to learned and professional societies such as the Royal Society of Chemistry, that mandate ORCID iDs as part of funding application and manuscript publishing process.

In a higher education landscape, PIDs for researchers such as ORCID have been raised through direct engagement with the researchers themselves, usually mediated through university libraries or research offices. Imperial College London, for instance, had one of the highest uptakes of ORCID iDs in the UK since an outreach programme[15] was launched as a joint collaboration between the university library and the research office.

ISNIs have been used to disambiguate individuals and organisations as contributors, distributors and rights holders of creative works since 2012. Assigning ISNIs does not rely on active contributors or claiming. ISNIs are therefore particularly useful as identifiers for individuals who may be long-since deceased or simply no longer actively creating outputs. ISNIs are assigned through disambiguation and matching by ISNI assignment agencies, based on biographical information and bibliographic information of works created. As well as uptake by libraries for identifying authors, contributors and distributors of scholarly outputs, ISNIs

[12] https://blog.datacite.org/pids-for-conferences/
[13] Dasler, R., Deane-Pratt, A., Lavasa, A., Rueda, L., & Dallmeier-Tiessen, S. (2017). Study Of Orcid Adoption Across Disciplines And Locations. Zenodo. https://doi.org/10.5281/zenodo.841777
[14] Armstrong, D., Haak, L., Meadows, A., & Stone, A. (2015). ORCID 2015 Survey Report (final) [Data set]. Figshare. https://doi.org/10.6084/m9.figshare.2008206.v1
[15] https://www.imperial.ac.uk/research-and-innovation/support-for-staff/scholarly-communication/orcid/project/

are finding wider information management uses, including recent adoption by YouTube for contributors to musical works on the platform[16].

While much of the technical infrastructure and approaches that underpin people PIDs is mature, the major maturity gap lies in uptake of PIDs by researchers across disciplines. Incentives for engagement by the "research society" need to be employed that transcend any cultural barriers. Integrations with publisher platforms and CRIS are currently minimal—the infrastructure exists yet only a fraction of the landscape is integrated. Relevant higher-level policies and governance could be introduced to ensure greater interoperability between different PID systems and thereby enable stronger links between stages, results and participants in the research lifecycle.

3.2.4 Organisations

The community currently has the ability to assign identifiers to published content and to individuals, however, a missing piece is a comprehensive, open, and accessible organisation identifier infrastructure that identifies affiliations and is independent of a particular organisation identifier service provider's business needs. While there are many examples in this space there is no single solution that meets the broad needs of the research community.

Three of the FREYA partner organisations, are actively discussing the development of an open organisation identifier registry (DataCite, Crossref and ORCID):

Crossref, DataCite, and ORCID came together to address the need for an open registry of organisation identifiers, which emerged from public reports and discussions that began at the Research Data Alliance Persistent Identifier Interest Group (PID-IG) meeting in 2015. Further progress was made in discussions at Coalition for Network Information CNI), FORCE11, and PIDapalooza meetings in 2016. Crossref, DataCite and ORCID announced the formation of an organisation Identifier Working Group[17] in 2017 to refine the structure, principles, and technology specifications for an open, independent, non-profit organisation identifier registry to facilitate the disambiguation of researcher affiliations. The Working Group developed principles for Registry governance and product requirements.

An outcome of the working group was a product definition for an organisation identifier registry. The scope of the organisation identifier registry focuses on affiliation and provides an open registry for the description of relationships between contributors, contributions, research sponsors, publishers, and employers. In January 2018 Crossref, DataCite and ORCID organized organisation ID Stakeholders meeting in Girona, Spain[18]. The purpose of the meeting was to gather stakeholders to discuss scope and governance model of a service to support an Open organisation Identifier registry. Work is underway to propose the organization and governance that can take this work forward.

As well as the developing organisation ID work, other PIDs are available to identify organisations in other spaces. For instance, ISNI currently provides identifiers for over 700,000 organisations in its registry. The ISNI IDs are maintained in a single database which is curated by a network of Members and Registration Agencies. It has one Agency dedicated to the specific curation of organisation identities, Ringgold[19], which manages around 500,000 ISNIs linked to its own proprietary Ringgold Org ID. Ringgold provides free ISNI lookup, API and download at isni.ringgold.com. A study by JISC-CASRAI[20] has recommended adoption of ISNI Org IDs for research management systems, advocating a role for national libraries. The British Library undertook a proof-of-concept project in 2016 assigning approximately 1000 ISNIs to organisations in the Research Councils UK database. The ISNI-International Authority (ISNI-IA) has begun working with a range

[16] http://www.isni.org/content/youtube-adopts-isni-id-artists-songwriters
[17] https://orcid.org/content/organisation-identifier-working-group
[18] https://orcid.org/content/2018-org-id-meeting
[19] https://www.ringgold.com/isni/
[20] The Jisc CASRAI-UK Organisational Identifiers Working Group's charter is described at http://jisccasraipilot.jiscinvolve.org/wp/working-groups/org-id/ (Archived at https://perma.cc/D66Y-MRXT).

of stakeholders to consult on optimizing ISNI Org IDs for the research sector. In a webinar held in May 2018 the ISNI-IA presented its roadmap for optimizing Org IDs highlighting the need to implement some technical changes and community requirements. The community requirement to meet was the set up of an advisory board to capture and represent stakeholder needs. The technical changes highlighted include:

- segmentation of Org IDs from the ISNI database with a searchable user interface;
- core metadata available under CC0 & in Linked Data formats;
- provide users with a periodically-updated file download of the Organisations Registry;
- API for retrieval of ISNI IDs and records with the ability to resolve an ISNI ID;
- online form for organisations to add metadata to their own records.

The main proposition of the ISNI-IA is that its Org IDs are maintained by a network of members and agencies so that the effort of creating and maintaining the identifiers is shared. The database and many of the required identifiers already exist but require ongoing curation, adoption by CRIS systems and consequently linking to their proprietary system IDs.

Other identifiers for organisations have so far been assigned to very specific kinds or organisations, in limited contexts. For instance DOIs for those that provide funding with the Crossref FunderID and funded organisations in the US who require a DUNS number (Data Universal Number System[21]). There are also the ARCHON code[22], an ID from the UK National Archives for archival institutions holding collections relating to the UK, and many countries apply an identification number to businesses[23] and charities[24].

3.2.5 Data

Research data (e.g. measurement data or survey data) enable researchers to verify results and pursue new research questions. It is critical to determine precisely which data were used to generate a certain outcome and to be able to access an exact dataset. PID systems are an essential component of data citation, as other metadata attributes do not unambiguously identify a particular item and cannot be used for a reliable location, retrieval or verification of research results. With ever growing data volumes and reprocessing of lower level data into new products, the use of unique and persistent identifier systems is increasing. With their ability to version dynamic datasets, PID systems offer solutions to meet the needs of an increasingly complex research landscape. As with publications, the PID type often depends on the exact data type and disciplinary background, and there is a great variety of established PID types for data. User communities in different disciplines need to decide on how to deploy PIDs going forward. As an example, for longitudinal studies the versioning of the resulting datasets is a topic of debate: Is a new PID required for each version of a dataset or should related versions of a dataset be archived with the same PID? Solutions may be different for social sciences versus life sciences.

According to a 2017 re3data analysis[25], the DOI system is the most common PID type implemented in research data repositories across all disciplines (20%), followed by the Handle system. A number of repositories provide other discipline or infrastructure specific PID types. Persistent Uniform Resource Locators (PURL), Uniform Resource Names (URN) and Archival Resource Key (ARK) are in mature use. An example of a discipline specific PID type is the European Case Law Identifier (ECLI) developed to facilitate accurate citation of judgments from European and national courts. Publishers are encouraged to provide a set of uniform metadata to improve discoverability in case law.

[21] https://www.grants.gov/applicants/organization-registration/step-1-obtain-duns-number.html
[22] https://www.wikidata.org/wiki/Property:P3642
[23] Usually from business registers such as Companies House in the UK (https://beta.companieshouse.gov.uk/) and the Australian Business Number (https://abr.business.gov.au/Home/About)
[24] For instance, in Canada (https://www.charitycentral.ca/rc4-what-charitable-registration-numberoe4-qu%E2%80%99est-ce-qu%E2%80%99un-num%C3%A9ro-d%E2%80%99enregistrement-%C3%A0-titre-d);
[25] Kindling, Maxi et al. (2017). The Landscape of Research Data Repositories in 2015: A re3data Analysis. D-Lib Magazine (23). https://doi.org/10.1045/march2017-kindling

A sobering statistic from the abovementioned re3data analysis, is that more than 65% of the indexed research data repositories do not yet provide a PID system and only 46 out of 1421 repositories have more than one PID system applied. Data citation is not currently in common practice, so even if a PID is assigned to a specific dataset, it is often not used for citation.

3.2.6 Data Repositories

Data repositories provide infrastructure to host research data. The Registry of Research Data Repositories, re3data[26], provides general information about repositories, but also additional data, such as information concerning supported persistent identifier systems.

Another registry of data repositories is FAIRsharing[27], initially focussed on the life sciences, but now with a broader scope. FAIRsharing also covers other resources, including standards and policies. Data repositories are also included in registries that describe all repositories, such as OpenDOAR.[28]

Although multiple registries describe data repositories, there is no commonly used persistent identifier assigned to them and no standard metadata. re3data uses DOIs and has a metadata schema, but that schema is not used by other research data registries. This is in contrast to journals and books which use the ISSN or ISBN, respectively. There is thus no unique identification of a data repository yet, and a recent editorial in the data journal Scientific Data highlighted the problem[29], but suggested standard naming instead of persistent identifiers as the solution. From the experience with people and organizations, this is approach is unlikely to solve the unique identification issue.

A persistent identifier for data repositories that is used across the data sharing community would solve a number of important use cases, and is not difficult to implement, given that data repositories are well-described entities, exist in relatively small numbers (re3data has 2,088 entries as of May 28, 2018), and the community is well-organized, e.g. via RDA.

3.2.7 Grants (Research Awards)

The sources of support for research is varied and the terms “grants” and “grant identifiers” are used inclusively here for “grants, endowments, secondments, loans, use of facilities/equipment and even crowd-funding”[30].

Funders have acknowledged a need for an “open & global grant identifier system” that would make the identification of research outputs associated with a grant more accurate, reduce the burden of tracking and be compatible with the existing in-house identifier schemes used by >15K+ funders. To this end, a pilot service is being planned[31] along the following lines: New grants awarded by life science funders including the Wellcome Trust, NIH, and MRC, will be assigned a unique identifier, likely a DOI, through a contract with Crossref. These DOIs will need to resolve to a publicly accessible repository such as the EuropePMC's Grants Finder Repository[32] which already archive grant data for 29 European funders. The plan is for the DOI to comprise the existing funder’s grant ID, prefixed with a funder ID—which can be taken from the Crossref Funder Registry[33].

ORCID includes grant identifiers in the funder activity section of ORCID records. Interactions with the funder community through the ORCID ORBIT Funder Working Group[34] make it clear there is a real need to

[26] https://www.re3data.org
[27] https://fairsharing.org
[28] http://www.opendoar.org
[29] What’s in a name? (2018). Scientific Data, 5, 180092. https://doi.org/10.1038/sdata.2018.92
[30] https://www.crossref.org/blog/global-persistent-identifiers-for-grants-awards-and-facilities/
[31] https://www.crossref.org/blog/wellcome-explains-the-benefits-of-developing-an-open-and-global-grant-identifier/
[32] http://europepmc.org/grantfinder
[33] https://www.crossref.org/services/funder-registry/
[34] https://orcid.org/about/community/working-group/funders

be able to track connections to grants. Currently this is limited to a funder specific local identifier due to the unavailability of a universally unique and resolvable identifier.

The suggested workflow for enabling output tracking is for funders to add funding activities which include grant identifiers to the ORCID records of the associated contributors, having captured authenticated ORCID iDs and permissions to update the records during the grant application process. When submitting articles, publishers can then present a list of relevant grants for the submitter to choose from. Chosen grant identifiers are then associated with the publishing metadata, hopefully using DOIs. Funders and others would then be able to easily find and locate the products of their grants.

The grants registration with ORCID can include grants-in-kind such as beamtime in large research facilities operating synchrotrons, neutron sources or X-ray lasers. This should allow impact measurements and advanced linking of various information entities similarly to the case of monetary grants. This view is promoted by ORCID User Facilities and Publications Working Group.[35]

3.2.8 Projects

Projects can be defined as targeted activities based using allocated resources such as budget, time and expertise.

The "project" is recognised as a key entity by CERIF (Common European Research Information Format)[36] which aims to formalize the connectivity between research entities. It is also a term used by several research information systems. There is currently no widely-adopted standard for the identification of projects. In Europe, local identifiers are used by project funders (as noted for grants, in the section above) and by Research Information Systems.

Projects are also identified by description fields and some of these can be considered to be identifiers. An example is EU funded projects—these can be identified by features including the name of the funding program, currently Horizon 2020. An overview of all H2020 projects [37] lists the following public information grant information for each: Record Control Number (RCN), project ID (grant agreement number), project acronym, project status, funding programme, topic, project title, project start date, project end date, project objective, project total cost, EC max contribution (commitment), call ID, funding scheme (type of action), coordinator, coordinator country, participants (ordered in a semi-colon separated list), participant countries (ordered in a semi-colon separated list).

An example of project information maintained by research information portals can be seen in the Dutch NARCIS system[38] that contains information of about 67,000 research projects. Here "project" entities are identified as "research". Research is defined as "Project descriptions of current and completed research projects e.g. per research discipline, programme, research school, organisation, researcher and index term". About 18,000 dissertation projects from all disciplines are part of the system. Each research project is identified by a unique code that starts with the string "OND" followed by a number.

Consultation with stakeholders could be undertaken to assess whether it is feasible to put effort into the development of a common standard for the persistent identification of projects.

Related to projects, DARIAH-EU, the pan-European research infrastructure consortium for arts and humanities, connects hundreds of scholars and dozens of research facilities in 17 European countries[39]. In

[35] https://orcid.org/content/user-facilities-and-publications-working-group;work by this group is also mentioned in the 'Equipment' section.
[36] https://www.eurocris.org/cerif/main-features-cerif)
[37] An overview of all H2020 projects with their public grant information can be found at: https://data.europa.eu/euodp/en/data/dataset/cordisH2020projects
[38] http://www.narcis.nl
[39] https://www.dariah.eu/

many cases the components of the DARIAH network are identified and related to each other by the use of PIDs.

RAiD is a persistent identifier for Research Activities. RAiD[40] places the research activity at the centre of the data life-cycle and is intended to have a many-to-many relationship with other PIDs. "Activity" is used here as an umbrella term and can be applied flexibly to identify projects, sub-projects, experiments, research programs and calibrations for any discipline or industry. The RAiD itself is a handle, with an attached metadata manifest which collects a timeline of related PIDs, such as DOIs, ORCID IDs, ISNI, GRID. RAiD was initially developed as part of the Australian Data LifeCycle Framework project, which is a nationwide strategy to connect research resources and activities created by NCRIS funding and co-investment[41] and supported by ARDC[42], AARNET[43] and AAF[44]. The DLCF[45] held meetings with university researchers, and research management professionals, Government groups, Research organisations and funding bodies across Australia and identified the need for a means to identify and track research through activity rather than by the researcher. RAiD was released in production in April 2017, and is integrated in a selection of Research Management Systems, with further test integrations underway in Australia and also in New Zealand. RAiD is currently progressing through the process to become an ISO standard[46], having been accepted as a "New Work Item" by Standards Australia. The current challenges for RAiD are: increasing integration so the utility of the PID can be demonstrated, and addressing lack of knowledge around RAiD and related PIDs.

Examples covered in the following sections indicate situations where the terms "projects", "experiments", "investigations", and "analyses" have a specific meaning for a discipline. Note that in the life sciences, these terms are used less specifically and at times interchangeably . e.g., The Human Genome *Project*, is described as an "international collaborative research *program"*[47], or The Global Ocean Sampling *Expedition* which was "a pilot sampling *project"*[48]. Likewise, "experiments", "investigations", and "analyses" are used interchangeably by laboratory or clinical researchers to refer to their research.

3.2.9 Experiments

In the field of High Energy Physics (HEP), "experiments" refer to either the hardware (e.g. particle detectors) of an experiment or a collaboration of people working on a specific research project with specific hardware. There are currently seven Large Hadron Collider (LHC) experiments at CERN. Each uses detectors to analyse particles produced by collisions in the accelerator. These experiments are run by global collaborations of scientists. Each experiment is distinct, and characterized by its detectors[49].

INSPIRE[50] is a High Energy Physics bibliographic search system curated by a consortium consisting of CERN, DESY, Fermilab, SLAC and IHEP. The information system indexes the High-Energy Physics Experiments Database[51] amongst other content. Here, experiments are categorized into collider experiments, fixed-target experiments, neutrino (flavor) experiments, dark matter search experiments, cosmic ray experiments, other rare-process/exotic experiments, accelerator test facility experiments, astronomy experiments and theory collaborations. The HEP Experiments Database contains basic metadata and links by URLs to experiment-specific content. INSPIRE does not assign PIDs for collaborations, but the

[40] https://www.raid.org.au/
[41] National Collaborative Research Infrastructure Strategy (NCRIS) - an Australian Government initiative
[42] Australian Research Data Commons - https://www.ands-nectar-rds.org.au/
[43] Australia's Academic and Research Network - https://www.aarnet.edu.au/
[44] Australian Access Federation - https://aaf.edu.au/
[45] The Data LifeCycle Framework - https://www.dlc.edu.au/about
[46] https://www.iso.org/standards.html
[47] https://www.sanger.ac.uk/news/view/human-genome-project-centres
[48] https://www.jcvi.org/global-ocean-sampling-expedition-gos
[49] https://home.cern/about/experiments
[50] http://www.inspirehep.net/
[51] http://inspirehep.net/collection/Experiments

collaboration members get indexed in HepNames where they can provide their ORCID iD. The reason for this is the dynamic nature of a collaboration, that can change on a daily basis.

An internal database-specific ID is used when citing the state of the hardware of experiments at CERN. The database is not open to the public, since the information it holds is deemed to be sensitive. A requirement of PIDs is that they resolve to openly accessible landing pages[52]. This would need to be addressed at CERN for PIDs to be assigned to these experiments.

3.2.10 Investigations

An "investigation" is a collective term used by large-scale research facilities such as a neutron source or synchrotron radiation source facilities to refer to several related experiments, with some of the experiments potentially used for instrument calibration and the rest for the purpose of data acquisition. Facilities typically assign unique identifiers to investigations; these are facility-specific rather than universally unique. "Investigation time" granted by facilities can be considered a non-monetary form of a research grant. Extending the analogy, an investigation ID can be deemed to be similar to a grant ID assigned by funding agencies. Some facilities assign persistent identifiers to investigations, using DOIs sourced through the Application Programming Interface of the DataCite service.

The DOI assigned to an investigation is associated with a landing web page supported by the facility on its own web server. The DOI can be assigned and the associated landing page created directly after the time-slot is granted to a visitor scientist and before the experiments commence. The landing page is then populated with metadata collected from the research proposal managed by a facility-specific proposal system. When the investigation commences and experimental data is collected, the landing page is supplied with a link to the data holdings, which may be restricted for an embargo period to the scientists who performed the experiments.

Notably, investigations have features in common with formal research publications - notably the systematic assignment of DOIs to investigations, their accompanying additional structured metadata and their inclusion in citation networks.[53]

3.2.11 Analyses (studies of data)

In High Energy Physics (HEP) analysis information is understood to be the combination of data and metadata. In this context, data means datasets, code, and/or results while metadata usually includes contextual information like the analysis name, contact persons or publications[54]. Jupyter Notebooks can be considered a type of Analyses. Git "gists" (see in Software section) are used sometimes as Jupyter Notebooks identifiers.

Preserving the entirety of an analysis for an experiment is crucial to reproducibility and transparency of the research process. This task remains challenging, as data resulting from LHC experiments are unique, costly, and complex - details such as the software, the underlying operating system platform and user analysis code used in a given physics analysis must be recorded in order to reproduce an experiment.

The CERN Analysis Preservation Framework (CAP)[55] functions as a central platform for all four LHC collaborations. It preserves information and tools for HEP analyses[56] and addresses the need for the long-

[52] https://support.crossref.org/hc/en-us/articles/214669863-Your-landing-page
[53] See figure 3 below; Bunakov, V. Investigation as a member of research discourse. In 16th All-Russian Conference "Digital Libraries: Advanced Methods and Technologies, Digital Collections", Dubna, Russia, 13-16 Oct 2014. CEUR Workshop Proceedings Vol-1297 (2014): 160-165. http://purl.org/net/epubs/work/12302226
[54] CERN Analysis Preservation Support Group (2017): CERN Analysis Preservation Documentation Release. URL: https://media.readthedocs.org/pdf/cernanalysispreservation/latest/cernanalysispreservation.pdf
[55] https://analysispreservation.cern.ch/ (only internally accessible)
[56] J. Cowten et al. (2015): Open Data and Data Analysis Preservation Services for LHC Experiments. Journal of Physics: Conference Series (664). DOI: 10.1088/1742-6596/664/3/032030

term preservation of all the digital assets and associated knowledge in the data analysis process. Recording high-level physics information such as physics object selection, relevant documentation and discussions is not currently required. The information structure on CAP represents the analysis workflow steps and reflects changes in content or in the workflow by versioning the analysis process and the underlying JSON schema. Access to content and the analysis chain is restricted to participants in the collaboration. Final plots, presentations and papers are typically made available to the public.

PID services for all analysis components have not yet been fully established, although CAP service functions are constantly being enriched and improved. Through CERN's participation in the THOR project, ORCID authentication and dynamic data citation was explored[57]. Within the context of FREYA, the implementation of PIDs in CAP is ongoing.

3.2.12 Software

Software code, crucial for computational methods, simulations and manipulating research data-sets, is increasingly being called upon to be made openly available for transparent review and reproducibility[58]. Further incentives to making software open, are that it can be properly cited and developers can receive credit for their work. Several initiatives currently facilitate software citation. One is the CERN-hosted online repository Zenodo[59], which allows source code from the popular software development site GitHub to be preserved and cited through its infrastructure by registering DOIs for research software[60]. Of the now more than 50k DOIs registered for software, more than 80% were registered via Zenodo[61].

In 2016 the Force11 Software Citation Principles were published, providing guidance for how software should be cited[62]. Among the recommendations is the use of persistent identifiers for software citation, also recommended in best practice recommendations[63]. PIDs for software need to address particular challenges associated with software, such as versioning. Another challenge is the variety of metadata standards used in the community. The Codemeta project[64] tries to address this challenge by providing a crosswalk table between these metadata standards, and codemeta support is added to an increasing number of services facilitating software citation[9].

Although software citation principles are broadly endorsed by the community and an increasing number of tools and services facilitate software citation, the practice of software citation is not yet centered around PIDs. Rather there are varied forms of software mentions in scholarly papers: provision of simple names in the full-text, URLs in footnotes, project names, websites, user manuals or prior publications listed in the references[65]. The Force11 Software Citation Implementation Working Group is working with the community to facilitate adoption of their recommendations.

---

[57] https://www.w3.org/2016/11/sdsvoc/SDSVoc16_paper_4

[58] V. Stodden, G. Peixuan and M. Zhaokun, "Toward Reproducible Computational Research: An Empirical Analysis of Data and Code Policy Adoption by Journals," PLoS One, vol. 8, no. 6, 2013, e67111. DOI: https://doi.org/10.1371/journal.pone.0067111

[59] https://zenodo.org/

[60] A. Purcell. Tool developed at CERN makes software citation easier (2014) Retrieved from http://cds.cern.ch/record/1998637

[61] Fenner, M., Katz, D. S., Nielsen, L. H., & Smith, A. (2018). DOI Registrations for Software. https://doi.org/10.5438/1nmy-9902

[62] Smith, A. M., Katz, D. S., & Niemeyer, K. E. (2016). Software citation principles. PeerJ Computer Science, 2, e86. https://doi.org/10.7717/peerj-cs.86

[63] Stodden, V. & Miguez, S., (2014). Best Practices for Computational Science: Software Infrastructure and Environments for Reproducible and Extensible Research. Journal of Open Research Software. 2(1), p.e21. DOI: http://doi.org/10.5334/jors.ay

[64] Boettiger, C. (2017, January). Codemeta: A Rosetta Stone for Software Metadata. figshare. https://doi.org/10.6084/m9.figshare.4490588

[65] Li, K. et al. (2017). How is R cited in research outputs? Structure, impacts, and citation standard. Journal of Informetrics 4 (11), pp. 989-1002. DOI: https://doi.org/10.1016/j.joi.2017.08.003

DOIs might be the preferred persistent identifier for citing research software in most communities, however there is another highly relevant identifier used to track changes during software code development (versioning): the "commit hash" of the version control system, e.g. git, that is based on Merkle trees. Hashes are derived directly from the software code, and (with some exceptions) globally unique. These hashes are used heavily by source code repositories such as GitHub, and by Software Heritage, an archive for software source code[66] that archives GitHub as well as the now defunct source code repositories Google Code and Microsoft CodePlex. Commit hashes are better suited for use than DOIs when versioning is key, and further work is needed to align the two identifiers.

Sha-1 (Secure Hash Algorithm 1)[67] and Git "gists"[68] can also be used as software identifiers that recognise the importance of versioning.

3.2.13 Computer simulations

Practices of using persistent identifiers for computer simulations can vary across disciplines. These differences are also due in part to the dual nature of computer simulations which can be considered either a subcategory of "software" (above) or in-silico experiments ("investigations" - see section below).

Materials Cloud[69] suggest URLs for citing certain simulations along with DOIs when a corresponding research paper is available.

The Centre for Environmental Data Analytics (CEDA)[70] use persistent URLs formed on a base of Universally Unique Identifiers (UUIDs). At minimum, the URL points to an abstract describing the computational model description; it can also point to any data resulting from the simulation or to DOIs of associated published research articles.

In HEP, high-energy simulations of particle collisions provide a detailed theoretical reference for the measurements performed at accelerators against which models of both known and new physics can be tested. The information can be highly granular, and refer to individual particles[71]. Published simulations (or simulated data) are released at CERN OpenData[72], where they are assigned DOIs. The metadata includes related derived data and relevant software. Unpublished simulated data is preserved in CERN Analysis Preservation (CAP). As discussed later, the process of integrating PIDs for items in CAP is ongoing.

3.2.14 License information for software

The current practise in software development is to use the URLs of the license text as the persistent identifier and include them into source files. This works well for most software licenses, because their URLs are stable (e.g. the URL of the Apache Software License, version 1.0, has not changed since 1999). With increasing numbers of websites switching to HTTPS, some licenses have multiple aliases (with HTTP and with HTTPS). Another problem for some license types (namely Creative Commons licenses) is that the URL can variably contain the language of the license text. Thus multiple URL variations exist for CC-BY licenses including a generic URL plus one for each license text language used. Those URLs should be normalized when used as identifiers (e.g., language removed from URL).

There are initiatives that maintain databases of Licenses. One example is Software Package Data Exchange (SPDX)[73] which aims to standardize the way that license information, including components, license and

[66] https://www.softwareheritage.org/
[67] https://en.wikipedia.org/wiki/SHA-1
[68] Git gists. https://www.labnol.org/internet/github-gist-tutorial/28499/; https://help.github.com/articles/about-gists/
[69] Materials Cloud portal. https://www.materialscloud.org
[70] Centre for Environmental Data Analytics. http://www.ceda.ac.uk/
[71] http://lhcathome.web.cern.ch/projects/test4theory/high-energy-physics-simulations
[72] http://opendata.cern.ch/
[73] https://spdx.org/

copyrights, is shared across the software supply chain. It contains a list of alias names or URLs that can be used for harmonizing those license URLs.

Adding a completely new identifier for licenses is not easily achievable with the current infrastructure, as License URLs are used in source code files. In summary, there is community consensus that there is no clear need for persistent identifiers for license information, or for additional work in this area.

3.2.15 Equipment

This section seeks to cover "resources" used in research, including instruments, devices, platforms, research spaces and storage facilities. A discussion of "samples", the subjects or outputs of research, is provided in a separate section of this report.

In the natural sciences, equipment such as vessels, platforms, buoys, sensors, sensor arrays or networks and other instrumentation are often central to data acquisition. While identifiers for vessels and platforms carrying instrumentation are relatively easily assigned, assigning identifiers for devices, instruments and sensors is more complex and these rarely bear any persistent identification other than inventory IDs in their owner's ledger. Large-scale and unique instrument descriptions can be published in the Journal of Large-Scale Research Facilities (JLSRF)[74] by the operating institution. Each article is assigned a DOI. An instrument upgrade that qualifies as a new instrument would warrant a further article in JLSRF with a new DOI.

Instrument performance depends heavily on lab/environmental conditions and maintenance procedures, including appropriate sensor calibrations. Instrument PIDs that include general information in the metadata such as the type, manufacturer, model, and manufacturer specifications for the instrument, are helpful as a first step. Missing still is information relating to replacement of hardware and software components or varying environmental and lab conditions that can influence measurements and/or detection rates to a large degree. Issuing a new PID after changes to an instrument is a debated solution for this but would erase the instrument's history. In turn, the instrument metadata could be separated into general metadata that is sent to the PID registration agency vs. specialized (and yet to be defined) metadata that could remain at the instrument database (i.e. the provider of landing pages). Separating metadata in this way is similar to what is in place currently for datasets and literature.

With few exceptions, the behaviour of recording metadata to reflect the state and status of the instrument at the time of data acquisition, is not in common practice. The need for well-defined guidelines for metadata is one focus of the RDA Group PIDINST-Working group on PIDs for instruments[75], a group convened by the JLSRF editorial team amongst others. The working group first met in March 2018 and is currently collecting case studies from various research centres with the primary aim to develop a common schema for metadata about instrument instances. In addition the working group will discuss the possibility of a schema for metadata that instrument database providers may consider providing on landing pages. Here, the working group also plans to issue best practices for the publication of such metadata in both human and machine readable format. The outputs of this group focus on the use of instrument PIDs and PID-associated metadata by machine agents; and complements the ongoing publication of "instrument articles" for human reading by JLSRF.

In marine research, the use of vessels, platforms, onboard and deployable instrumentation in highly varying environments is very extensive. Several initiatives have recently been installed to try to tackle the inclusion of equipment in data metadata, supporting the full flow of sensor observations to archives. The US ***"Rolling Deck to Repository (R2R)"***[76], is an exemplary initiative here, recording and providing digital data generated by environmental sensor systems permanently installed on research vessels. These R2R "Cruise-level" metadata records also include type and model of each instrument system along with file format and release

[74] Journal of large-scale research facilities (JLSRF). https://jlsrf.org/
[75] https://www.rd-alliance.org/group/persistent-identification-instruments/case-statement/persistent-identification-instruments
[76] http://www.rvdata.us/about/products

status. An instrument identifier (DOI) is issued on completion of metadata records that links these to the data files generated by the instrument. Individual instruments can also be tracked via instrument serial numbers, if available and provided. However, non-permanent equipment is not yet included. Such equipment can make up a large share of the instrumentation onboard research vessels. Also missing from the metadata is the state and status of the instrument at the time of data acquisition and there is no option to account for an equipment state or status change.

The *Alfred-Wegener Institute (AWI)*, which coordinates German polar research, provides further solutions for equipment identification and accounting in research. It has recently (2015) initiated the Sensor Information System infrastructure[77] to support the flow of sensor observation to archives. They built a cost-effective and generic framework, the "**Observations to Archive (O2A**)", which complies with OGC standards, ensuring interoperability in an international context (e.g. SOS/SWE, WPS, WMS WFS,..). Each sensor is described following SensorML data model standards and data is fed to an SOS interface, so that the sensor can be monitored in real or near real-time. Here too, gaps persist in reflecting equipment status and state at different points in time.

The R2R initiative is fortunate to employ the intensive oversight of its vessel operators and staff to acquire and maintain the records of ship-bound equipment. It is not clear how other stakeholders will provide the equivalent detail of information for their equipment. Uptake could be facilitated if equipment users were provided (by the AWI, for instance) with the guidelines and tools to enable equipment identification and accounting.

In the life sciences, research institutions such as the European Molecular Biology Laboratory (EMBL) offer a range of **core technology facilities**[78] for high-end microscopy and image analysis, functional genomics analyses, flow cytometry, to name a few. In response to calls by many such institutions, to track usage of individual facilities and the research outputs resulting from their technical support, several working groups have been convened to understand the requirements and possibilities for developing the PID infrastructure around facilities. The User Facilities and Publications Working Group[79] grew out of initial community conversations in 2017. The group, led by ORCID, have worked to define 'research resources' and assess how to use identifiers to record usage. Earlier this year ORCID announced plans to create a new 'Research Resources' section in the ORCID record[80]. The group is currently planning community engagement around defined pilot projects.

3.2.16 Archival/Storage facilities

These include storage facilities for and descriptions of historical records, as well as archival finding aids, described as tools that facilitate discovery of information within a collection of records. Finding aids usually include a description of the scope of the collection, biographical and historical information related to the collection, and restrictions on use of or access to the materials[81]

For documentation used in archival finding aids, the Encoded Archival Description (EAD)[82] is of eminent importance. The standard contains a large number of elements such as the "Record" identifier: (with the tag <recordid>) which "is used for recording a unique identifier for the EAD instance." The institution assigning the identifier ensures uniqueness of the <recordid> value within the archival entity under its control. A globally unique identifier may be added such as HTTP URI, DOI, PURL, or UUID), or used in combination with the <agencycode>, which is a required element of the <maintenanceagency>"[83].

[77] https://sensor.awi.de/
[78] https://www.embl.de/services/core_facilities/index.html
[79] https://orcid.org/content/user-facilities-and-publications-working-group
[80] https://orcid.org/blog/2018/04/10/acknowledging-research-resources-new-orcid-data-model
[81] https://www2.archivists.org/glossary/terms/f/finding-aid
[82] https://www.loc.gov/ead/
[83] https://www.loc.gov/ead/EAD3taglib/tl_ead3.pdf, page 337 of EAD Tag Library.

The International Council on Archives (ICA) promotes the use of metadata standards for archival descriptions, ISAD[84] and authority records, ISAR[85] , but PIDs are relevant for creation of these metadata records.

Identifiers for archival entities are currently being assigned and include URI, DOI, and UUID.

3.2.17 Research stations

Identifiers for research stations have been under discussion but are not yet in common use. One initiative involved the organisation of Biological Field Stations (OBFS)[86] and the University of California's Natural Reserve System, who are looking to cite field stations. Their field sites are located around the globe, are federal, state, or private property, government funded, and/or managed by a host institution. Researchers, not formally affiliated with the sites, travel to the sites (often with researchers from other institutions) to conduct their work. The field stations have a need to track the output (publications, data, etc) associated with a field station, but there is no clear or standard way to cite field stations. Discussions have lead to a proposal to create individual, citable "description" with associated persistent identifiers for each field station (more generically called a "site"). Key here is a description of a field site (location, characteristics, etc). A few years ago BioScience published an article, The Way Forward for Biological Field Stations[87], that clearly articulates the need.

DataCite offers an example of how a particular organisation uses identifiers for research sites: The International Federation of Digital Seismograph Networks (FDSN)[88] is a global organisation. Its membership comprises groups responsible for the installation and maintenance of seismographs within their geographic borders or globally. FDSN is truly global - not dominated by any one country or group, and includes members from all continents. Most members of the FDSN operate stations that are confined to their national boundaries but several FDSN members operate stations well outside their borders. FDSN have created a DOI for each member node[89] in the network with metadata that describes the particular site. FDSN encourages their researchers to associate the DOI with any outputs (data, software, publication, etc) from the network node. This allows all research generated at a particular site to be grouped together and provides the means for the site to get credit.

There is scope for very fine levels of granularity for research stations. One example of this comes from long-term field experiments. At the Rothamsted Research facility at Harpenden in the UK, the Park Grass experiment[90] is a field trial that has been ongoing since 1856. A uniform field was subjected to various soil enrichment treatments to investigate how each treatment affected hay yield. Over the past 160 years each treatment area has subsequently been subdivided, with different treatments applied. This has resulted in a field facility with plots that have been subject to different conditions that could impact all subsequent data from each location within the experiment going forward. Soil and crop samples are available from the facility, as well as the yield data going back to year one. For provenance, and for accurate interpretation and analysis of results, it is important for researchers to be able to identify the geolocation of each plot and relate it back to its historical treatments.

A related need emerging in the cultural sector is identification of storage and display locations for physical artefacts. Being able to identify exactly which room, wall or even wall area that an artefact was displayed on, contextualises research into an artifact's condition and physical composition. As well as contextualising

[84] https://www.ica.org/en/isadg-general-international-standard-archival-description-second-edition
[85] https://www.ica.org/en/isaar-cpf-international-standard-archival-authority-record-corporate-bodies-persons-and-families-2nd
[86] http://www.obfs.org/
[87] http://bioscience.oxfordjournals.org/content/65/2/123.full.pdf
[88] http://www.fdsn.org/
[89] http://www.fdsn.org/networks/
[90] http://www.era.rothamsted.ac.uk/Park

a physical research object, such information is also useful for ephemeral items of research such as exhibitions.

The need for identifying such a granular aspects of a facility like this will necessitate the metadata to include related PIDs in order to contextualise the entity. Beyond that, it's likely that most other metadata will be very subject-specific. In the case of the Park Grass Experiment, that would include geolocation and historical treatment regimes with dates.

3.2.18 Samples

In geological and biological sciences, physical samples of biological or geological origin can be stored and undergo multiple analyses by various researchers with different scopes. In this way, the scientific outcome of a collected sample can increase. Furthermore, identification of the physical samples enhance reproducibility of research allowing for quality control and tests of data from previous analyses. For these purposes, linking physical samples with data and publications is essential.

Physical Samples from the natural environment can be assigned an ID in form of the **International Geo Sample Number (IGSN).** The IGSN is a unique alphanumeric code, which can be assigned through the registration of physical samples. IGSN was developed by the System for Earth Sample Registration (SESAR)[91], which also issue IGSNs along with various other Allocating Agents following the same IGSN rules and regulations. IGSNs can be assigned to samples from a broad range of origin: rock, mineral, and fossil specimens, dredges, cores (rock, sediment, ice), fluid samples (seawater, river or lake water, hydrothermal fluids, pore water), drill holes and wells, soil pedons, macro- and micro-biological samples and more. IGSNs are Handles that can be resolved to landing pages describing the sample.

To trace scientific outcome and enhance reproducibility of a given sample, it is very valuable to be able to identify the physical sample that was used for a given study, and if preserved, where this sample can be located. Hence, cross referencing ID for physical samples (IGSN) with both research and data publications is essential. Elsevier and Copernicus earth science journals have recently implemented the use of IGSNs. Implementation of IGSNs is also recommended by the Coalition for Publishing Data in the Earth and Space Sciences (COPDESS). PANGAEA[92] has also long been publishing data and metadata that include IGSNs, thereby allowing for the data to be traced back to specific physical samples.

The cross linking between PIDs for physical samples and data is not always a trivial matter. Even though IGSNs are Handles there is no guarantee of a digital registration of the samples with the appropriate metadata assigned. Further difficulties revolve around PIDs for physical samples that follow sample ID systems other than IGSN, which may be specific to scientific disciplines, countries or institution. PANGAEA identifies an IGSN and links it to the metadata of the dataset. Currently, IGSNs are not linked to DataCite and Schema.org metadata, but various initiatives are working towards expanding the implementation of IGSNs in international data management.

In the life sciences, the Resource Identification Initiative[93] has set up a portal[94] whereby persistent identifiers known as **Research resource identifiers (RRIDs)** can be assigned to antibodies, cell lines, model organisms as well as some tools. RRIDs use established community identifiers where they exist. In these cases, identifiers are prefixed with " RRID: ", followed by a second tag that indicates the source authority that provided it (e.g. "AB" for the Antibody Registry, "CVCL" for the Cellosaurus, "RGD" for Rat Genome Database, "SCR" for the scicrunch registry of tools).

[91] http://www.geosamples.org
[92] PANGAEA is a data publisher for earth and environmental science; https://www.pangaea.de/
[93] Bandrowski et al. (2015) The Resource Identification Initiative: A cultural shift in publishing. Co-Published: Journal of Comparative Neurology [10.1002/cne.23913], Brain and Behavior [10.1002/brb3.417], F1000 Research [10.12688/f1000research.6555.2], and Neuroinformatics [10.1007/s12021-015-9284-3].
[94] https://scicrunch.org/resources

RRID-labelled entities could be considered to be "samples" in the true sense in that they are the end result (analytes) of a study; or they could be considered to be "equipment" i.e. reagents used to generate results and end-products. Currently >2000 publications cite RRIDs[95], a contribution that aids transparency and reproducibility of findings. However, issues such as granularity of the samples being identified, limit global adoption.

3.2.19 Cultural artefacts

Where research within the humanities relies on physical cultural artefacts, it is important to be able to identify that artefact precisely. However, the kinds of "cultural artefacts" that could be an item of research range from a painting, to an archaeological find such as a piece of jewellery, through to buildings and aircraft. Within this broad range, there are a number of factors to be considered around persistent identification, such as:

- political and cultural sensitivities about persistent identification as a potential statement of ownership and authority;
- transfer of individual items between organisations and the implications that would have for managing a PID;
- both the above relate to additional concerns with the return of spoliated (i.e. damaged) items.

Galleries, Libraries, Archives and Museums (GLAM) holding cultural artefacts that are the subject of research do have their own internal identifiers and accession numbers. For example, the British Museum have museum registration numbers such as "Af2004,04.1"[96] (referring to a Ghanaian garment) and the Smithsonian Museums' Inventory number such as "A19510007000"(their Bell X-1 aircraft)[97]. Artefacts themselves may also give rise to samples (e.g. a paint sample[98] is from NG1259 as cited by Gent et al (2014)[99]). Beyond citation needs, a persistent link is required here for provenance, allowing a link between an artefact and its samples, as well as the data from those samples and any digital representations of the artefact itself.

Reports of archaeological research are stored in Europe in several national repositories. In the Netherlands the archeological datasets are stored in the EASY system[100] and they are uniquely identified by a DOI and URN. The EU project Ariadne developed a portal for access to archaeological resources located across Europe[101]. All records in the portal are identified by a unique "Ariadne ID".

There is currently no widely-adopted standard for the identification of cultural artefacts. Schemes that do exist such as MuseumID[102], the Cultural Objects Name Authority[103] or the PICHE project (Persistent Identifiers for Cultural Heritage Entities)[104] run by the German National Library are not in use beyond a handful of organisations. MuseumID and PICHE are based on URNs, but neither have wide adoption. Meanwhile, some GLAM organisations continue to build systems to meet their own needs. London's National Gallery are creating their own PID system to enable identification of their works of art (for example https://data.ng-london.org.uk/resource/000-03Y9-0000 for NG1259) and also identification of

[95] https://elifesciences.org/inside-elife/ff683ecc/rrids-how-did-we-get-here-and-where-are-we-going
[96] https://www.nationalgallery.org.uk/media/23860/volume35essay2reynoldstech.pdf
[97] https://www.si.edu/object/nasm_A19510007000
[98] https://research.ng-london.org.uk/projects/technical-bulletin/vol-35/tb35-technical%20essay/images/M1381s9_vis_20x
[99] https://www.nationalgallery.org.uk/media/23860/volume35essay2reynoldstech.pdf
[100] http://easy.dans.knaw.nl
[101] http://portal.ariadne-infrastructure.eu/
[102] http://museumid.net/documentation
[103] http://www.getty.edu/cona/CONAFullSubject.aspx?subid=700000157; http://www.getty.edu/research/tools/vocabularies/cona/
[104] http://www.kim-forum.org/EN/Wir/Projekte/Laufend/piche.html

locations for those artworks within the gallery ("006-000N-0000" is Room 34), artists ("001-03F6-0000" is Sir Joshua Reynolds) and keywords ("00A-0045-0000" is Cupids).

3.2.20 Historical and mythical personae

The need to identify persons or "person-like entities" in historical research is not the same as name identifiers for the researchers themselves. The identifier would not necessarily be tied to the entity itself, but rather a particular interpretation or manifestation of the entity. References to "Cleopatra" could refer to the historical person, or the Shakespearian character or, or many of the other representations of that person over time. Similar to ancient world places, there are a number of resources that identify ancient persons or person-like entities. The Standards for Networking Ancient Prosopographies: Data and Relations in Greco-Roman Names project (SNAP:DRGN)[105] aims to bring these together, providing persistent identifiers for historical entities for use in linked data[106]. Similar to the historical place names, metadata includes names and name variants, dates and sources[107].

3.2.21 Temporal periods and historical places

Within the arts and humanities, spans of time may have slightly different definitions with the "Bronze Age" having different start and end years depending not only on the location in question, but the opinion of the writer. It is therefore vital to specify which definition of a temporal period is being mentioned, particularly when reusing, combing and visualising data.

One attempt to rationalise and identify interpretations of historic periods is PeriodO. Each of their period definitions is assigned an ARK[108] to allow humanities data to be directly linked to a specific time span. Metadata for each item includes a title, source, start and end dates for the period and notes on the origin and record. Temporal periods can thus be considered to be "samples": identifiers point to the source of the data under scrutiny, with metadata that describes the period of time.

These rationale are similar for the identification of historical places. While geographic coordinates give an explicit location, place names have borders and time spans that are open to interpretation, making it important to define what is meant when a certain place name is used to define data. There are a large number of examples of gazetteers that enable translation of historical places into contemporary geolocations, many referencing locations at a fine level of granularity, giving a high level of detail for local interest. Examples of larger-scale gazetteers for historical place names are Pleiades[109], which covers the ancient world (namely Greek, Roman, the Ancient Near East, Byzantine, Celtic and Early Medieval worlds), Trismegistos[110], also covering the ancient world and the Getty Thesaurus of Geographic Names[111]. All of these provide URIs for place names[112] and common metadata fields include geographic coordinates, place name and name variants, location types (e.g. "Deserted settlement" or "settlement"), and sources or references. Projects such as Pelagios[113] aim to bring disparate resources for the ancient world together in one resource[114].

A specific example concerns the boundaries of cities and villages in the Netherlands which have been registered since 1812. These are uniquely identified by two codes: "Amsterdamse code" and "CBS code".

[105] www.geonames.org/
[106] https://kclpure.kcl.ac.uk/portal/files/56550302/a44_lawrence.pdf
[107] See: https://snapdrgn.net/cookbook and https://doi.org/10.1145/2786451.2786496 (via: https://kclpure.kcl.ac.uk/portal/files/56550302/a44_lawrence.pdf)
[108] e.g. http://n2t.net/ark:/99152/p06c6g3mrpb or http://n2t.net/ark:/99152/p0pf7xr4x4z
[109] https://pleiades.stoa.org/
[110] https://www.trismegistos.org/geo/
[111] http://www.getty.edu/research/tools/vocabularies/tgn/
[112] https://pleiades.stoa.org/places/79739, www.trismegistos.org/geo/14800 and http://vocab.getty.edu/page/tgn/7011881
[113] http://commons.pelagios.org/
[114] http://oro.open.ac.uk/43658/1/2014_Isaksen_Barker_etal_Pelagios_WebSci.pdf

For current names of geographical locations, the Geonames database[115] is used. These identifiers can be linked to coordinates with specific "time stamps". Unique codes and identifiers are available that "fix" the boundaries and names of geographical locations to a given year. Based on these connections the extension of boundaries can be expressed, thereby making it possible to demonstrate, for instance, how Amsterdam annexed surrounding villages over the course of time[116] or how city boundaries changed over time[117].

3.2.22 Study registrations

Clinical trials

For clinical sciences, a research study protocol refers to a research plan and is therefore broader than the individual technique or recipe. A research study protocol is required for all studies involving human subjects whether observational or interventional. It is a "set of rules" or an overview of the research to be undertaken for a particular study including the rationale for the study, the objectives, the approach (which likely involves several techniques), and participants involved in the study (investigators as well as research subjects). Patient identifying information is carefully managed/excluded and not accessible to the public. Ahead of commencing with the clinical study or trial, these protocols need to be approved by the institution's regulatory body and the information registered (in a repository such as Clinical trials.gov[118] which was launched in Feb 2000 and supplies a registry identifier e.g. NCT01802372 for each deposition).

The International Committee of Medical Journal Editors recommend that journal editors only accept for publication clinical trial reports where the trial was registered within the WHO International Clinical Trials Registry Platform (ICTRP)[119]. The World Health Organisation in turn requires that a trial registration must have an identifier in order to be fully registered[120], but there is no consistent approach to these identifiers at present. The WHO provides comprehensive guidelines for metadata that should accompany trial registrations.The ISCRTN (International Standard Randomised Controlled Trial Number) is one such registry and curated database that issues DOIs. The metadata required for clinical trials submitted using CrossRef or DataCite is less comprehensive in that it excludes method-specific information such as details of the "study design" or "intervention". The ICMJE recommendation has been in place since 2005, and although links between articles and trial registrations are being made based on identifiers, they are not often linked using the persistent resolvable identifier when one is available[121].

Non-clinical study registration

While study registration is best established for clinical studies, it also gaining ground in other disciplines such social and economic research. Best practices for metadata registration and publication are less well established, but work is ongoing to develop this area. Identifier use for study registration is generally more comparable across disciplines. ISRCTN[122] and the Open Science Framework registry both provide DOIs (e.g. https://doi.org/10.17605/osf.io/xemzv) while most others like ClinicalTrials.gov, use accession numbers, e.g. the American Economic Association"s registry[123] and the Registry for International Development Impact Evaluations[124].

[115] www.geonames.org/
[116] http://www.gemeentegeschiedenis.nl/cbscode/0363
[117] http://www.gemeentegeschiedenis.nl/gemeentenaam/Amsterdam
[118] https://clinicaltrials.gov/
[119] http://www.icmje.org/recommendations/browse/publishing-and-editorial-issues/clinical-trial-registration.html
[120] http://www.who.int/ictrp/network/trds/en/
[121] for example see https://doi.org/10.1186/s12893-015-0061-x
[122] https://www.isrctn.com/
[123] e.g. https://www.socialscienceregistry.org/trials/352
[124] http://ridie.3ieimpact.org/index.php?r=search/detailView&id=374

3.2.23 Data Management Plans (DMPs)

Akin to a clinical research study protocol, the DMP relates to the data generated during a research project: describing what will be generated, how it will be generated; how will it be preserved and shared, noting any restrictions on sharing. Funders who are seeking to maximise research outputs associated with their grants, require these of fundees. Indeed provision of DMPs is a mandate of the EU-funded Horizon 2020 research and innovation programme[125]. While the requirement for DMPs provides the impetus to researchers to think carefully about data management, there are currently no formal repositories or portals for management plans per se.

There are several data management planning tools available to the community: the *DMPTool*[126] from the California Digital Library, *DMPonline*[127] from the Digital Curation Centre in the UK and *EasyDMP* [128]. These offer resources and tools for creating, maintaining and exporting data management plans (e.g. lists of funders' DMP requirements, examples of DMPs, software, etc).

There has been limited discussion about DMP transparency: the DMPtool makes PDFs of DMPs publicly available and the journal RIO Journal[129], is rare in that it accepts DMPs for publication[130].The latter have associated DOIs.

3.2.24 Workflows

A workflow is "the sequence of processes through which a piece of work passes from initiation to completion"[131]. A study registration or data management plan would fit this definition, but rather than a word-based workflow, what we intend to discuss here is a computational workflow, as commonly used for bioinformatics in the life sciences. In this case a workflow is considered to be "a series of tools and dataset actions that run in sequence as a batch operation" according to Galaxy (http://usegalaxy.org), a publicly available, web-based platform via which researchers can make their computational biomedical workflows and research accessible, reproducible, and transparent. Other widely used bioinformatics workflow tools include Apache Taverna[132] and Unipro UGENE[133].

Galaxy requires users to register a "public name" which is then used as an identifier to generate addresses (URIs?) for information that a user subsequently shares publicly.

Taverna generates URIs for workflow definitions, workflow run information, and produced data values but does not provide a repository to store this information. Their identifiers include using a "taverna" in the prefix and they add UUIDS to build these URIs.

A UK-based initiative called myExperiment[134] launched in 2007, captured the spirit of an online community of users and developers. myExperiment provides a repository for deposition and discovery of bioinformatics workflows and encourages sharing and derivative reuse. myExperiment have trialled

[125] http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf#page=10
[126] https://dmptool.org/; https://dmptool.org/public_plans
[127] http://www.dcc.ac.uk/dmponline
[128] https://easydmp.eudat.eu/plan/ . This online tool is developed by SIGMA2 (the national infrastructure for computational science in Norway) in collaboration with EUDAT and OpenAIRE.
[129] https://riojournal.com/about#WhatCanIPublish
[130] https://riojournal.com/browse_journal_articles.php?form_name=filter_articles&sortby=0&journal_id=17&search_in_=0&section_type%5B%5D=231
[131] Oxford English Dictionary
[132] https://taverna.incubator.apache.org/
[133] http://ugene.net/
[134] Goble, C.A., Bhagat, J., Aleksejevs, S., Cruickshank, D., Michaelides, D., Newman, D., Borkum, M., Bechhofer, S., Roos, M., Li, P., and De Roure, D.: myExperiment: a repository and social network for the sharing of bioinformatics workflows, Nucl. Acids Res., 2010. **doi:10.1093/nar/gkq429**; http://www.myexperiment.org/

assigning DOIs to items[135] although none after 2015. It partners with a selection of projects, including Galaxy and Taverna platforms, and its usage statistics indicate that it has attracted >10K registered members and hosts 3900 workflows to date.

3.2.25 Protocols

In the life sciences, a research protocol is the description of a specific method or technique used in a study e.g. a recipe for tissue culture media or the methodology for digesting DNA. A protocol may comprise a well established technique that may have been may be adapted for the specific use case. Alternatively, the technique may be newly devised. The details of a protocol, including the internal controls put in place to avoid confounding results, are required to establish the validity of the end results and to be able to reproduce the results. Repositories such as Protocols.io[136] allow for deposition of individual protocols, and their derivatives. Protocols are private by default, can be shared semi-privately with a research community of choice, or can be made public - under a CC-BY license. DOIs can be reserved via CrossRef prior to formal publication in order to cite the protocol within the "materials and methods" section of the research output. When accepted for publication, the publication will be listed in the protocols.io entry.

There are no PID services that enable researchers to claim protocols to their ORCIDs (identifiers for researchers).

[135] For examples see https://search.datacite.org/works?query=myexperiment

[136] Teytelman L, Stoliartchouk A, Kindler L, Hurwitz BL (2016) Protocols.io: Virtual Communities for Protocol Development and Discussion. PLoS Biol 14(8): e1002538. https://doi.org/10.1371/journal.pbio.1002538; https://www.protocols.io/

4 Concluding observations and musings

This report captures a snapshot of the PID landscape as understood by the partners involved in project FREYA[137]. It lists research entities (including publications, data, researchers, grants, equipment, etc.) used by the research community; and indicates a sense of the uptake and the PID services available that link the identifiers assigned to each entity.

Related entities have been **grouped into themes** for this report and the rationale behind the themes is as follows. In different disciplines, research entities are referred to in ways that make sense to the communities working on those disciplines, but by abstracting the sense and purpose of the research entity, it is possible to see that entities across different disciplines, with different names, are actually the same thing (in an abstract sense). A good example of this is a given time period (arts and humanities) when compared to physical samples (life sciences, geosciences). These may be the subject of (resource) or output of a scholarly work. Defining both, unambiguously, and then specifying which definition has been used in a scholarly work is key to both disciplines.

There will likely be further discussion on semantics and broader themes in which entities have been grouped in this report (types of projects, equipment, samples, study registrations). Community feedback is welcomed to enhance the accuracy and inclusiveness of this summary.

There are a few broad observations about **PID usage** that highlight the complexity of the landscape.

Table 1 lists what seems to be a host of different identifiers. However, there are actually not that many identifier types, if compact identifiers are considered one type. Compact identifiers comprise any local unique identifier with a prefix that is "repository identifying"[138].Examples include amongst many, GO:0003214 (a Gene Ontology record), ISBN:9780141392127 (International Standard Book Number that identifies Charles Dickens' Oliver Twist).

Some identifiers are used for a specific entity exclusively, such as the ORCID iD, which is used only for people. On the other hand, DOIs are used for more than one research entity type, originally for identifying journal publications, now data, moving to software and beyond. This highlights the need at a high level to specify the nature of the entity being identified by a PID type. This will make it clear for both human and machine readers what type of entity a DOI references—a computational workflow or a paper or a dataset. Furthermore, different research cultures in different disciplines lead to different behaviours and/or uptake rates for specific PID types, for example, the life scientists predominantly use compact IDs for data whereas environmental scientists use DOIs (see data published by PANGAEA). In the life sciences, it is also possible for the same compact identifier to resolve to several different sites; for example, in the case of the Protein Data Bank (PDB) the same identifier refers to the same protein structure dataset, but this dataset may be found at any of the nodes of the international collaboration, WorldwidePDB (wwPDB).

An entity such as people or publications can be referred to by several equivalent PID types. For example, ISNI/ORCID iDs for individual researchers; and PMID/PMC/DOI for publications. There are historical and functional reasons for this.

- First, repositories need to manage their own records, not just resolve (point to the location where a specific entry can be found). Taking the example of PMID/PMC/DOI for *publications*: a journal

[137] Members of the following partner institutions have contributed to this document: EMBL/EBI, The British Library, CERN, Datacite, DANS, STFC, PANGEAE, MARUM, ORCID, CROSSREF. Thanks to members of ANDS and PLOS for internal review.

[138] Wimalaratne SM , Juty N , Kunze J , Janée G , McMurry JA , Beard N , Jimenez R , Grethe JS ,Hermjakob H , Martone ME , Clark T . Uniform resolution of compact identifiers for biomedical data. Sci Data [08 May 2018, 5:180029] https://europepmc.org/abstract/MED/29737976

article (with a DOI that is allocated by a publisher and metadata registered with the DOI registry, Crossref) may also have its abstract indexed by PubMed if it is biomedical, and will be assigned a PMID (PubMed reference number) by the National Library of Medicine (USA). The PMID refers to the metadata record in PubMed. If the full text version of the same journal article (already identifiable via a DOI and PMID) is indexed in EuropePMC/PMC (an archive of full-text journal articles) then it will be assigned a PMC identifier, which refers to the full text version in EuropePMC/PMC. The PMID/PMC/DOI identifiers are equivalent in that they refer conceptually to the same article, but the specific instances are different. Another example of identifier type reflecting functionality is discussed in the *researcher* section of the document: ORCID iDs are assigned via a process that involves the researcher applying for his/her ORCID iD; whereas ISNIs do not rely on active claiming by a researcher and can be assigned for example to deceased scholars. In these cases, mapping across different identifiers is common practice.

- Second, resources may also operate mixed models of identifier assignment. For example in the PRIDE proteomics database, both a DOI and an accession number are allocated to a given proteomics study. As each study may contain 2 or 3 million data points, only accession numbers are used to uniquely identify the constituent data points.

Entities such as *protocols, workflows, data management plans* or *publications* and *investigations* (as used by large-scale research facilities), may be closely related enough to benefit from shared approaches for development of new PID services. It is important that the PID community apply lessons learned from technologies and governance of small or locally-operating initiatives when looking to scale interoperability more broadly.

A draft **maturity matrix for current PID services** is presented in Table 1. PIDs for *researchers*, *publications* and *data* are in widespread use, with uptake actively growing, and "mature" supporting services in place. However, there are entities described in this report that currently do not use PIDs but would benefit, or where a variety of disparate PIDs are in use in local systems, or where a mature PID system is used for a new entity type, but with limited uptake. In Table 1 these are listed with PID infrastructure noted as "emerging" or "immature". In particular there is growing interest from a number of stakeholders to further address the need for open and global identifier systems for entities such as *organisations*, *grants, software, research facilities and conferences*. Wider PID adoption and interoperability would enable a number of use cases to be addressed, such as better monitoring of how research resources have been used or understanding the impact of research outputs.

Moving forward, FREYA partners are collating user stories from stakeholders; these will be mapped to the landscape described in this report so as to identify opportunities for partners to grow the PID ecosystem and to prioritise where best to take action. Activities envisaged include consultation with key community stakeholders on PID implementations (e.g. via existing working groups), advocacy to promote existing PID uptake, and prototyping novel PID services. In the longer term, we expect that a growing number of entities will be assigned PIDs and therefore avail themselves to become integrated into the emerging growing PID graph.

Annex A: Abbreviations

ARK Archival Resource Key

BibCode Bibliographic Codes

CRIS Current Research Information Systems

DAI Digital Author Identifier

DOI Digital Object Identifier

EOSC European Open Science Cloud

FORCE11 The Future of Research Communications and eScholarship (working group convened in 2011)

ID Identifier

IGSN International Geo Sample Number

ISBN International Standard Book Number: An international ID for published books

ISNI International Standard Name Identifiers

ISNI-IA ISNI-International Authority

ISSN International Standard Serial Number

OCI Open Citation Identifier

OGC Open Geospatial Consortium

ORCID Open Researcher and Contributor ID

PID persistent identifier

PMID PubMed ID

PURL Persistent Uniform Resource Locators

RAiD Research Activity identifier

RDA Research Data Alliance

RDA PID-IG RDA Persistent Identifier Interest Group

RRID Research Resource ID

SHA-1 Secure Hash Algorithm 1

SOS OGC Sensor Observation Service

SWE OGC Sensor Web Enablement

UUID Universally Unique Identifiers

URI Uniform Resource Identifier

URL Uniform Resource Locator

URN Uniform Resource Name

VIAF Virtual International Authority Files

WFS Web feature Service

WMS Web Map Service

WPS Web Processing Service