D6.1
A system design for data identifier and citation services for environmental RIs projects to prepare an ENVRIPLUS strategy to negotiate with external organisations

WORK PACKAGE 6 – INTER-RI DATA IDENTIFICATION AND CITATION SERVICES

LEADING BENEFICIARY: LUND UNIVERSITY

Author(s):	Beneficiary/Institution
Margareta Hellström (lead), Monica Lassi and Alex Vermeulen	Lund University (LU)
Robert Huber, Markus Stocker	Universität Bremen (UniHB)
Frank Toussaint	Deutsches Klimarechenzentrum GmbH (DKRZ)
Malcolm Atkinson	The University of Edinburgh (UEDIN)
Markus Fiebig	Norsk institutt for luftforskning (NILU)

Accepted by: Zhiming Zhao (Theme 2 leader)

Deliverable type: REPORT

Dissemination level: PUBLIC

Deliverable due date: 30.12.2016/M20

Actual Date of Submission: 31.01.2017/M21

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant

ABSTRACT

Environmental research infrastructures are often built on a large number of distributed observational or experimental sites, run by hundreds of scientists and technicians, financially supported and administrated by a large number of institutions. If these data are shared under an open access policy it becomes therefore very important to acknowledge the data sources and their providers. There is also a strong need for common data citation tracking systems that allow data providers to identify downstream usage of their data so as to demonstrate their importance and show the impact to stakeholders and the public. Work Package 6 highlights identification and citation in environmental RIs, reviews available technologies and develops common services for these operations. This deliverable presents a suggested common system design for Identification and Citation, as well as an outline for negotiations and discussions with publishers and other actors in the scholarly data management and curation world. In addition, the report summarises the associated technological needs and requirements of the ENVRIplus partners.

Reviewer(s):

Reviewer(s):	**Institution**
Markus Stocker (internal)	Universität Bremen (UniHB)
Malcolm Atkinson (internal)	University of Edinburgh (UEDIN)
Daan Broeder (external)	Royal Netherlands Academy of Arts and Sciences (KNAW) and EUDAT
Marko Scholze (external)	Lund University (not involved in ENVRIplus)

Document history:

Date	**Version**
14.11.2016	Initial draft for WP6 comments
13.12.2016	Second draft sent for WP6 comments
18.01.2017	Version sent to reviewers
30.01.2017	Corrected version, sent for project review & approval
30.01.2017	Accepted by Theme 2 leader (Zhiming Zhao)

DOCUMENT AMENDMENT PROCEDURE

Amendments, comments and suggestions should be sent to the editors (Margareta Hellström margareta.hellstrom@nateko.lu.se and Alex Vermeulen alex.vermeulen@icos-ri.eu).

TERMINOLOGY

A complete project glossary is provided online at the ENVRIplus website, see: https://envriplus.manageprojects.com/s/text-documents/LFCMXHHCwS5hh.

PROJECT SUMMARY

ENVRIplus is a Horizon 2020 project bringing together Environmental and Earth System Research Infrastructures, projects and networks together with technical specialist partners to create a more coherent, interdisciplinary and interoperable cluster of Environmental Research Infrastructures across Europe. It is driven by three overarching goals: 1) promoting cross-fertilization between infrastructures, 2) implementing innovative concepts and devices across RIs, and 3) facilitating research and innovation in the field of environment for an increasing number of users outside the RIs.

ENVRIplus aligns its activities to a core strategic plan where sharing multi-disciplinary expertise will be most effective. The project aims to improve Earth observation monitoring systems and strategies, including actions to improve harmonization and innovation, and generate common solutions to many shared information technology and data related challenges. It also seeks to harmonise policies for access and provide strategies for knowledge transfer amongst RIs. ENVRIplus develops guidelines to enhance transdisciplinary use of data and data-products supported by applied use-cases involving RIs from different domains. The project coordinates actions to improve communication and cooperation, addressing Environmental RIs at all levels, from management to end-users, implementing RI-staff exchange programs, generating material for RI personnel, and proposing common strategic developments and actions for enhancing services to users and evaluating the socio-economic impacts.

ENVRIplus is expected to facilitate structuration and improve quality of services offered both within single RIs and at the pan-RI level. It promotes efficient and multi-disciplinary research offering new opportunities to users, new tools to RI managers and new communication strategies for environmental RI communities. The resulting solutions, services and other project outcomes are made available to all environmental RI initiatives, thus contributing to the development of a coherent European RI ecosystem.

TABLE OF CONTENTS

ABSTRACT....2
DOCUMENT AMENDMENT PROCEDURE....2
TERMINOLOGY....2
PROJECT SUMMARY....3
TABLE OF CONTENTS....4
1 ABOUT WORK PACKAGE 6....7
2 MOTIVATION....8
2.1 Background — Identification....8
2.2 Background — Citation....9
2.3 The Research Data Alliance (RDA) perspective....10
2.3.1 The Global Digital Object Cloud and the PID-centric model of data management....11
2.4 The FAIR, FORCE11, and CODATA-ICSTI Principles....13
2.4.1 The FORCE11 Principles....13
2.4.2 The FAIR Guiding Principles....14
2.4.3 The CODATA-ICSTI data citation principles....14
3 RI REQUIREMENTS FOR IDENTIFICATION & CITATION....15
3.1 Data identification requirements....15
3.1.1 Data citation requirements....15
3.1.2 Assessment of the I&C requirements....16
4 COMPONENTS OF PID IMPLEMENTATION, DATA PUBLISHING AND CITATION....16
4.1 The Persistent Identifier zoo....16
4.1.1 The Handle System (HS)....17
4.1.2 Digital Object Identifier (DOI)....17
4.1.3 Uniform Resource Name (URN)....17
4.1.4 Persistent URL (PURL)....18
4.1.5 Archival Resource Key (ARK)....18
4.1.6 Life Science Identifier (LSID)....18
4.1.7 “Cool” Uniform Resource Identifiers (CoolURIs)....18
4.2 Identifiers for non-data entities....19
4.3 Landing pages....20
4.4 “Session PIDs” vs “Citation PIDs”....21
4.5 Data publishing....22
4.6 Citation analysis....24
5 TECHNOLOGIES FOR IDENTIFICATION & CITATION....24
5.1 Technology review....24
5.1.1 Two-to-five year analysis of state of the art and trends....25
5.1.2 Details underpinning the two-to-five year analysis....26
5.1.3 A longer term horizon....31
5.2 Summary of analysis highlighting implications and issues....32
6 MAPPING THE PID PROVIDER, PUBLISHER & INDEXER LANDSCAPE....33

6.1 PID providers 33
6.1.1 ePIC — the European Persistent Identifier Consortium 33
6.1.2 DataCite 33
6.1.3 ORCID — Open Researcher and Contributor ID 34
6.1.4 ISNI — International Standard Name Identifier 34
6.1.5 Others 35
6.2 Associations and societies for scholarly publishers 36
6.2.1 Associations and societies for scholarly publishers 36
6.3 Indexing agencies and services 36
6.3.1 Crossref 36
6.3.2 Web of Science and Data Citation Index 37
6.3.3 Mendeley 37
6.3.4 Altmetrics 37
6.3.5 Making Data Count 38
6.3.6 Bibliographic databases 38
6.3.7 Discovery services 38
6.4 Global research data-oriented initiatives 38
6.4.1 Research Data Alliance (RDA) 38
6.4.2 CODATA 39
7 IDENTIFICATION & CITATION FOR RESEARCH DATA: WORK IN PROGRESS 40
7.1 Parallel activities by ENVRIplus partners 40
7.1.1 University of Bremen (PANGAEA) 40
7.1.2 DKRZ (CMIP6 project) 41
7.1.3 Marine RIs 42
7.1.4 ENVRIplus implementation cases 44
7.2 Other ongoing projects & initiatives 45
7.2.1 Project THOR 45
7.2.2 Scholix – Scholarly Link Exchange 46
7.2.3 OpenAIRE 46
8 A SUGGESTED SYSTEM DESIGN FOR ENVRIPLUS 47
8.1 Introduction to our approach 47
8.2 Data identification and citation in the ENVRI Reference Model 48
8.2.1 Data Identification in the RM 48
8.2.2 Data citation in the RM 49
8.3 Data Identification & Citation in practice — recommendations to RIs 50
8.3.1 Identification best practices for RIs 50
8.3.2 Citation best practices for RIs 54
8.4 Technologies, services and tools 56
9 NEGOTIATIONS WITH PUBLISHERS AND OTHERS 56
9.1 The case for negotiations 56
9.2 Discussion partners 57
9.3 Wish-list for services 57
9.3.1 Generic identifier minting services serving individual RIs 58
9.3.2 Dynamic data, including support for versioning 58

9.3.3 Support for inclusion of sub-setting information in citations....58
9.3.4 Management of data collections....59
9.3.5 Sustainable data typing services....59
9.4 The negotiation timeline....59
10 CONCLUSIONS....59
11 IMPACT ON PROJECT....60
12 IMPACT ON STAKEHOLDERS....60
REFERENCES....61

1 ABOUT WORK PACKAGE 6

The main goals of ENVRIplus Work Package 6 — Inter RI data identification and citation services — is to design and implement data tracing and citation functionalities in the environmental Research Infrastructures (RIs), and to develop tools for RI partners, if not otherwise available. The overarching objective is therefore to improve the efficiency of data identification and citation by providing convenient, effective and interoperable identifier management and citation services. This WP is also concerned with mapping out already existing service developments and studies related to data identification and citation that are being performed in the framework of European and global organisations and RI consortia in call EINFRA-7-2014,

As stated in the WP6 Description of Work, environmental research infrastructures are often built on a large number of distributed observational or experimental sites, run by hundreds of scientists and technicians, financially supported and administrated by a large number of institutions. If these data are shared under an open access policy it becomes therefore very important to acknowledge the data sources and their providers. There is also a strong need for common data citation tracking systems that allow data providers to identify downstream usage of their data so as to demonstrate their importance and show the impact to stakeholders and the public. This work package highlights identification and citation in environmental RIs, reviews available technologies and develops common services for these operations.

WP6 is organised around one task, T6.1. This aims at implementing common policy models for persistent identifiers for publishing and citing data. Moreover, the services for assigning and handling identifiers and for retrieving data content based on identifiers will also be provided. This task will build on existing approaches and current activities undertaken by ENVRIplus partners, and — if needed — synchronise with developments that arise from up-coming studies and projects from both service providers (ePIC, DataCite, EUDAT) and initiatives based in research organizations (THOR, OpenAIRE). It will be, furthermore, operated in close cooperation with existing initiatives (e.g. Research Data Alliance, ICSU WDS) and will elaborate a common data citation solution for the involved RIs.

This first deliverable from WP6, entitled “A system design for data identifier and citation services for environmental RIs projects to prepare an ENVRIPLUS strategy to negotiate with external organisations”, addresses the following points outlined in the Description of Work:

- Collect and promote the needs and priorities related to data identification & citation of environmental RIs in the global context, and present these two initiatives targeting pan-European Digital Identifier e-infrastructures as well as global initiatives such as the Belmont Forum and the Research Data Alliance. The discussions should lead to a widely accepted and supported model for a range of PID management-related topics.
- Perform an analysis of the latest statuses of these existing technologies and business models now used by PID service providers, publishers and data hosting organizations, and transfer the best and most common solutions to the RIs.
- Support negotiations on collaboration and contracts with important publishers. Publishers are an important partner in developing a functioning system of data citation. There are different models already available (journals for data description, direct citation via DOI, and data citation systems). Since environmental RIs provide large amounts of important data they can efficiently support respective negotiations.
- Define policy models for persistent identifiers, publishing and citation, and investigate how well these are represented in currently existing PID services such as those offered by ePIC and EUDAT.

2 MOTIVATION

In this chapter, we introduce some key concepts related to Identification and Citation. We also put the subject and the report into context.

2.1 Background — Identification

A number of approaches have been applied to solve the questions of how to unambiguously identify digital research data objects [Duerr 2011]. Traditionally, researchers have relied on their own internal identifier systems, such as encoding identification information into filenames and file catalogue structures, but this is neither comprehensible to others, nor sustainable over time and space [Stehouwer 2014]. Instead, data object identifiers should be unique "labels", registered in a central registry database that contains relevant basic metadata about the object, including a pointer to the location where the object can be found as well as basic information about the object itself. Exactly which metadata should be stored in the identifier registry, and in which format, is a topic under discussion, see e.g., [Weigel 2014]. Many environmental observational datasets pose a special challenge in that they are not reproducible, which means that also fixity information (checksums or even "content fingerprints") should be tied to the identifier [Socha 2013].

As a complement to the registry database, a lookup, or resolver, service is essential. When supplied with a valid identifier, the service should either return the associated metadata, or -- as is more common -- redirect to the supplied resource location. This can either be a direct link to the persistently identified object itself (e.g. a path to a file stored on a disk), or to a so-called landing page. The latter typically contains some basic metadata about the object, as well as information about how to access it.

[Duerr 2011] provide a comprehensive summary of the pros and cons of different identifier schemes, and also assess nine persistent identifier technologies and systems. Based on a combination of technical value, user value and archive value, DOIs (Digital Object Identifiers provided by DataCite) scored highest for overall functionality, followed by general handles (as provided by e.g., CNRI and DONA) and ARKs (Archive Resource Keys). DOIs have the advantage of being well-known to the scientific community via their use for scholarly publications, and this has contributed to their successful application to e.g., geoscience datasets over the last decade [Klump 2015]. General Handle PIDs have up to now mostly been used to enable referencing of data objects in the pre-publication steps [Schwardmann 2015] of the research data life cycle (illustrated in **Figure 1**). They could however in principle equally well be applied to finalised "publishable" data.

Persistent identifiers systems are also available for research-related resources other than digital data & metadata, articles and reports—it is now possible to register many other objects,

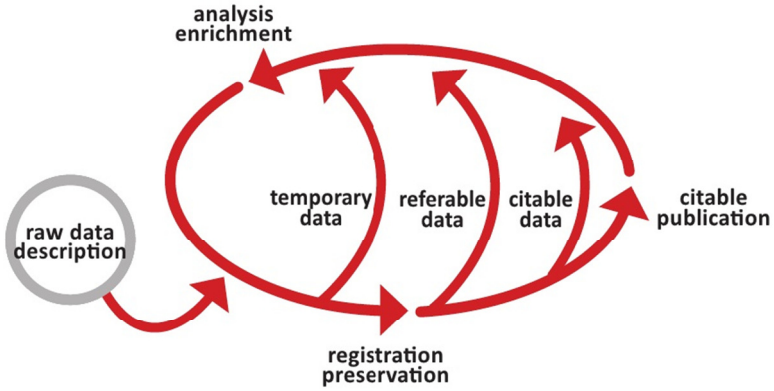

FIGURE 1. THE RESEARCH DATA LIFE CYCLE. DATA INTENSIVE RESEARCH IS HIGHLY COLLABORATIVE. ALLOCATING PERSISTENT IDENTIFIERS TO DATA OBJECTS SUPPORTS (RE-)USE AND SHARING OF DATA ALSO IN EARLY STAGES OF THE RESEARCH LIFE CYCLE. AFTER [SCHWARDMANN 2015].

including physical samples (IGSN), software, workflow processing methods— and of course also people and organisations (ORCID, ISNI). In the expanding "open data world", PIDs are an essential tool for establishing clear links between all entities involved in or connected with any given research project [Dodds 2014].

2.2 Background — Citation

The FORCE11 Joint Declaration of Data Citation Principles (JDDCP) [FORCE11 2014a] states that in analogy to articles, reports and other written scholarly work, also data should be considered as legitimate, citable products of research. (There is however currently an on-going discussion as to whether datasets are truly "published" if they haven't undergone a standardised quality control or peer-review, see e.g., [Parsons 2010].) Thus, any claims in scholarly literature that rely on data must include a corresponding citation, giving credit and legal attribution to the data producers, as well as facilitating the identification of, access to and verification of the used data (subsets). A generic workflow for data citation is presented in **Figure 2**. The workflow consists of a citation from a document to a dataset, a landing page in the repository where the dataset is stored, and the dataset itself.

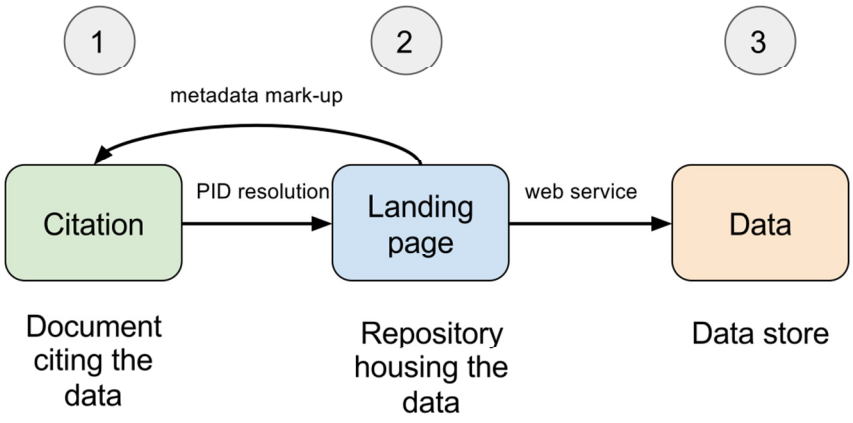

FIGURE 2. A GENERIC DATA CITATION WORKFLOW. SOURCE: [FORCE11 2016].

Data citation methods must be flexible, which implies some variability in standards and practices across different scientific communities [FORCE11 2014a]. However, to support interoperability and facilitate interpretation, the citation should preferably contain a number of metadata elements that make the dataset discoverable, including author, title, publisher, publication date, resource type, edition, version, feature name and location. Especially important, the data citation should include a persistent method of identification that is globally unique and contains the resource location as well as (links to) all other pertinent information that makes it human and machine actionable. In some (sensitive) cases, it may also be desirable to add fixity information such as a checksum or even a "content fingerprint" in the actual citation text [Socha 2013].

Finding standards for citing subsets of potentially very large and complex datasets poses a special problem, as outlined by [Huber 2013], as e.g., granularity, formats and parameter names can differ widely across disciplines. Another very important issue concerns how to unambiguously refer to the state and contents of a dynamic dataset that may be variable with time, e.g., because new data are being added (open-ended time series) or corrections introduced (applying new calibrations or evaluation algorithms) [Rauber 2015, Rauber 2016].

Both these topics are of special importance for environmental research today.

A number of surveys have indicated that the perceived lack of proper attribution of data is a major reason for the hesitancy felt by many researchers to share their data openly [Uhlir 2012], [Socha 2013], [Gallagher 2015]. This attitude also extends to allowing their data to be

incorporated into larger data collections, as it is often not possible to perform micro-attribution – i.e., to trace back the provenance of an extracted subset (that was actually used in an analysis) to the individual provider – through the currently used data citation practices.

2.3 The Research Data Alliance (RDA) perspective

The Research Data Alliance (see **Chapter 6.4.1**) interest group on Data Fabric issues (https://www.rd-alliance.org/group/data-fabric-ig.html) has chosen persistent identifiers as one of its focus areas (called "bundles"). As part of their on-going study, the group has collected and summarised statements and views from various research data management stakeholders and experts on a number of related topics, including the following about PID registries and PID usage:

PID registries

In order to be trustworthy and accepted on a global level, a system for minting, managing and resolving PIDs should be maintained by a dedicated and reliable team that operates under the oversight of a non-profit organisation which is itself governed by international boards. The system must be based on a transparent sustainable business model, and the activities be subject to regular quality assessments by external parties. Furthermore, the PID registry must be built on a redundant and secure architecture, based on open standards, and be accessible through an openly documented API optimally supporting accepted data models. To cope with the rapidly expanding need and use for PIDs across many more fields than just research, the system should be designed and constructed to support a huge address space (comparable or even larger than IPv6). Concerning functionality, the PID record should be able to store not just a bare minimum of attributes for the digital objects (bit stream location or landing page URL, owner, date) but also information about the objects' context (metadata, fixity, rights information, data type, etc.).

PID usage

Concerning the best practices of identifier usage, a PID needs to be requested as early as possible in the data object's life cycle, since at least at the time of curation at a trustworthy repository a PID record needs to be available. PIDs are not only useful for individual objects; they can also be associated with collections which can consist of a large number of digital entities. This allows the level of granularity at which PIDs will be assigned to be left to the communities and repositories. Metadata descriptions associated with a data object need to contain the PID of the corresponding object in order to allow the two to be connected. Ideally, the metadata can be given its own PID — e.g. pointing to a metadata file or to a query to a cataloguing system. Conversely, the PID record of the data object contains the metadata PID to ensure at all times that the data object's context can be retrieved. (This is in some ways similar, but not identical, to the reverse DNS mechanism used for internet IP addresses and domain names.) Finally, the PID record should offer the possibility to define an expiration date for the digital object. However, even for digital objects that have been deleted, it is important that the PID records should persist, and be updated with a notice about the deletion and, if possible, a pointer to the respective metadata records.

With the basis in the Data Fabric group's work, the European section of RDA (RDA Europe) brought together a number of international experts on PIDs with representatives of research infrastructures and librarians at the workshop "Views about PID systems", held at Garching, Germany in 2016 [Beck 2016]. ENVRIplus was represented with a talk on how to ensure that data producers are given appropriate credit when their work is included in data collections. One of the aims of the workshop was to identify areas of agreement and disagreement between the participants. Another concerned the possibilities to identify a given PID technology that could be used for all kinds of objects. Here, the Handle system tended to be favoured; even though other technologies could work equally well for e.g. research data, they would be less well suited for dealing with high volumes of objects (such as items produced by the manufacturing industry, or individual sensors of large detector arrays).

Some key conclusions of the discussions and presentations during the two-day event include [Beck 2016]:

- Proper PID usage and support will become key for competitiveness in science and industry.
- PIDs need to be used by all parties dealing with data professionally to make full use of advanced opportunities. A PID centric approach to data management, access and use will open the way towards new and comprehensive approaches to data handling and finally to a Global Digital Object Cloud as a generic, non-proprietary virtualisation layer.
- International and national steps need to be taken urgently to offer a sustainable, structured and mature PID service landscape based on quality assessed service providers to all interested parties. Only such a structured and massive approach will prevent ending up with unresolvable PID zombies.
- PIDs are becoming essential across sectors and communities for different application scenarios and efforts need to be taken to offer services across these sectors and communities.
- Setting up and maintaining trustworthy repositories is key for a structured data landscape guaranteeing access to data and its accompanying metadata.
- We need to design the required mechanisms (for facilitating automatic data processing) and build the needed tools now with high urgency.
- We urgently need to come to a structured and integrated domain of Handle Service Providers.
- Service providers need to ensure that these two interoperable domains are part of one integrated landscape of rich services.
- The PID centric approaches that are key to managing the “data tsunami” (brought on by “big data”) require simple and clear messages for the users.

2.3.1 The Global Digital Object Cloud and the PID-centric model of data management

Research in almost all disciplines is becoming more and more data-driven, and the current volumes and complexity of collected and otherwise generated data are increasing at ever higher rates. In addition, data are being shared and reused in ways, and in interdisciplinary contexts, that could not have been imagined only a few years ago. This “data tsunami” cannot be efficiently handled by current technologies that are built on domain-specific management concepts. (There are objections to the term “data tsunami”, as it has very negative connotations. Indeed, “data bonanza” may be much more appropriate!)

Within the framework of the Research Data Alliance (RDA) Data Fabric interest group (https://www.rd-alliance.org/group/data-fabric-ig.html), work is underway to identify and define core components of data management & handling — as used throughout the data life cycle — with the aim to enhancing interoperability in a way that benefits both human and machine consumption of data. Data-intensive research poses a number of challenges, including the need for good cataloguing systems that support fast discovery and retrieval of datasets based on complex searches. Machine-actionable workflow engines must be able to autonomously decide if a given search result actually contains useful information, if access is authorised and can be achieved under a suitable data license, and how to optimally access and process the bit stream.

A Digital Object (DO) can be defined as a bit sequence that has associated quality metadata describing it and a unique and persistent identifier [Weigel 2016]. The persistent identifier record must contain basic information such as the pointer to the DO’s location, but it can also be augmented with other usable attributes [Lannom 2016]. Indeed, to allow for any useful interlinking between DOs and other (digital) objects and resources, it is not enough to simply store the object’s location in the PID record, but this should also contain the pointer to the full metadata description — and reciprocally, the PID record of the metadata object (MO) should

also include the pointer to the DO. Other metadata that can be included in the DO's PID record include its data type(s), links to the workflow that produced it, recommended software for visualizing its contents, licensing information and much more. But note that as far as it is possible, the PID record should contain *pointers* (ideally in the form of HTTP PIDs, themselves resolvable on the Web and actionable) to where the related information is stored (in a trustworthy way) — and not the information itself — as illustrated in **Figure 3**. (The two alternatives are referred to as "call by reference" and "call by value", respectively.) In this way, the PIDs act as the glue that sticks the data fabric together, allowing proper interpretation and reuse as they persistently store all the necessary actionable references to the locations of the bit sequences [Lannom 2016].

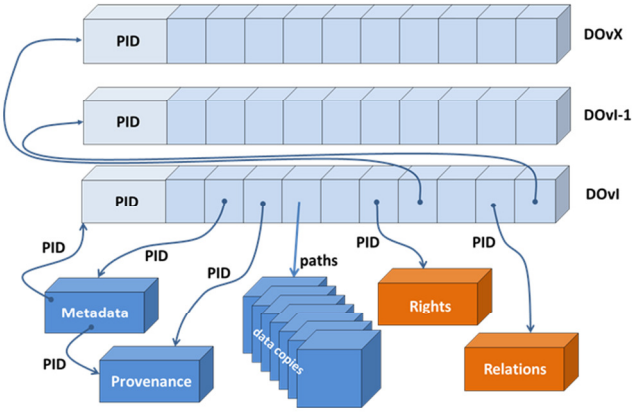

FIGURE 3. USING INFORMATION IN PID REGISTRY ENTRIES TO CONNECT A DIGITAL DATA OBJECT WITH ITS DESCRIPTIVE METADATA, PROVENANCE, RIGHTS & LICENSING INFORMATION AND ITS RELATIONS TO OTHER DATA OBJECTS. FROM [WEIGEL 2016].

A domain containing registered DOs of the type just described can be referred to as a Global Digital Object Cloud (GDOC). GDOCs are by definition based on the ideas of a Digital Object Architecture (DOA) and are fully compliant with the FAIR principles. **Figure 4** below illustrates the cloud concept schematically. In this model, the cloud of Digital Objects (second panel from the left) comprise a virtualization layer on top of network resources and services — in an analogy to how files and databases can be seen as virtualization layers on top of raw computer storage. Each of the DOs is persistently and uniquely identified and can thus be unambiguously referenced, which allows it to be operated on by end users and processes. In addition, the DOs are described and typed by their metadata, ideally using links to definitions stored in Data Type Registries (DTRs).

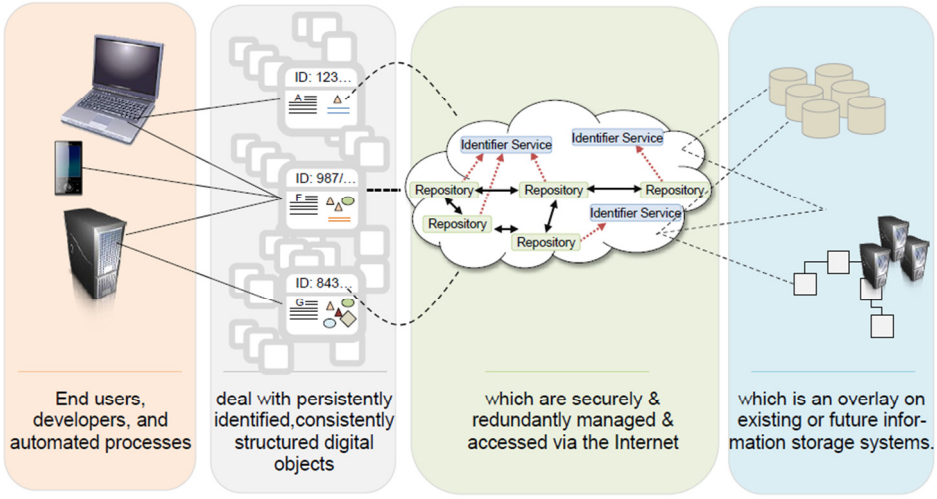

FIGURE 4. THE DIGITAL OBJECT CLOUD AND ITS CONNECTIONS TO USERS AND BACKEND STORAGE SYSTEMS. FROM [WEIGEL 2016].

However, this Linked Data-based approach will require powerful PID resolution services that are able to support the reconstruction of a digital DO of the type shown in Figure 3 that is sufficient for its processing by e.g. a machine-actionable workflow engine. Depending on the underlying scientific method, it may be first required to collate the necessary characteristics of the DO by computing the transitive closure of a large graph via many, potentially slow, visits to the resolution service. To optimise these kinds of operations, it will be necessary to balance data object granularity with the choice of where to store object metadata.

2.4 The FAIR, FORCE11, and CODATA-ICSTI Principles

In this section, we briefly summarise basic principles, statements and recommendations related to the management of digital research data objects that have been issued in the last couple of years from international expert groups including FORCE11 and CODATA-ICSTI.

2.4.1 The FORCE11 Principles

The Joint Declaration of Data Citation Principles (JDDCP) have been widely accepted by data repositories and publishing organisations [Fenner 2016a]. The principles of the JDDCP are: Importance, Credit and attribution, Evidence, Unique identification, Access, Persistence, Specificity and verifiability, and Interoperability and flexibility [FORCE11 2014a]. A data citation roadmap for scholarly data repositories has been developed by the Repositories Early Adopters Expert Group, which belongs to the Data Citation Implementation Pilot project of FORCE11 and BioCADDIE (https://biocaddie.org/), with members from DataCite and a wide range of data repositories spanning disciplines such as medicine, social sciences and biology [Fenner 2016a]. The roadmap suggests a set of data citation practices with the aim to meet the principles, and facilitate data citation for publishers and data repositories alike.

The roadmap's recommendations are made on three levels in order to help data repositories to prioritise: required – needed to meet the Joint Declaration of Data Citation Principles; recommended – to support publishing workflows together with publishers; and optional – to support data citation by data repositories [Fenner 2016a]. The recommendations are:

Required:

1. All datasets intended for citation must have a globally unique persistent identifier that can be expressed as an unambiguous URL.
2. Persistent identifiers for datasets must support multiple levels of granularity, where appropriate.
3. This persistent identifier expressed as URL must resolve to a landing page specific for that dataset.
4. The persistent identifier must be embedded in the landing page in machine-readable format.
5. The repository must provide documentation and support for data citation.

Recommended:

6. The landing page should include metadata required for citation, and ideally also metadata helping with discovery, in human-readable and machine-readable format.
7. The machine-readable metadata should use schema.org mark-up in JSON-LD format.
8. Metadata should be made available via HTML meta tags to facilitate use by reference managers.

Optional

9. Content negotiation [Wikipedia 2017a] for schema.org/JSON-LD and other content types may be supported so that the persistent identifier expressed as URL resolves directly to machine-readable metadata.
10. HTTP link headers may be supported to advertise content negotiation options
11. Metadata may be made available for download in BibTeX or other standard bibliographic formats.

2.4.2 The FAIR Guiding Principles

The FAIR guiding principles for data, developed by FORCE11, describe how to make published data Findable, Accessible, Interoperable, and Reusable for potential users [FORCE11 2014b, Wilkinson 2016]. Findability means making the data possible to find by potential users, e.g. by describing the data with rich metadata. Accessibility means that the data and metadata should be usable in formats that are understandable by humans and machines, e.g. by adding machine-actionable PIDs. Interoperability pertains to using a metadata scheme that is open and well-defined. Reusability means that the metadata are verifiable, machine-readable and can be used to make proper citations [FORCE11 2014b].

To be Findable:

F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:

A1. (meta)data are retrievable by their identifier using a standardised communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

To be Interoperable:

I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

To be Reusable:

R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

By following the FAIR principles, the "7-R's" of data can be fulfilled, i.e. that data are reusable, repurposable, repeatable, reproducible, replayable, referenceable, and respectful [Bechhofer 2013].

2.4.3 The CODATA-ICSTI data citation principles

The thorough review of current and emerging data citation practices in a wide range of disciplines by the CODATA-ICSTI Task Group on Data Citation Standards and Practices [Socha 2013] resulted in a set of principles for data citation: status, attribution, persistence, access, discovery, provenance, granularity, verifiability, standards, and flexibility.

1. The Status Principle: Data citations should be accorded the same importance in the scholarly record as the citation of other objects.

2. The Attribution Principle: Citations should facilitate giving scholarly credit and legal attribution to all parties responsible for those data.

3. The Persistence Principle: Citations should be as durable as the cited objects.

4. The Access Principle: Citations should facilitate access both to the data themselves and to such associated metadata and documentation as are necessary for both humans and machines to make informed use of the referenced data.

5. The Discovery Principle: Citations should support the discovery of data and their documentation.

6. The Provenance Principle: Citations should facilitate the establishment of provenance of data.

7. The Granularity Principle: Citations should support the finest-grained description necessary to identify the data.

8. The Verifiability Principle: Citations should contain information sufficient to identify the data unambiguously.

9. The Metadata Standards Principle: Citations should employ widely accepted metadata standards.

10. The Flexibility Principle: Citation methods should be sufficiently flexible to accommodate the variant practices among communities but should not differ so much that they compromise interoperability of data across communities.

3 RI REQUIREMENTS FOR IDENTIFICATION & CITATION

As part of the activities of the ENVRIplus Work Package 5 (REF), a review was undertaken to map out and assess the ICT and data management requirements, issues and opportunities of the Environmental Research Infrastructures (RIs) engaged in ENVRIplus. Data Identification & Citation was one of twelve topics that were investigated. In the following, we summarise the most important outcomes for Identification & Citation that were reported in Deliverable D5.1 [Atkinson 2016]. More information, including a listing of the questions that were asked, as well as the detailed responses of the participating RIs, is also available at the ENVRI Technology Review wiki site (https://wiki.envri.eu/display/EC/Identification+and+citation+requirements).

3.1 Data identification requirements

The survey found a large diversity between RIs regarding their practices. Most are applying file-based storage for their data, rather than database technologies, which suggests that it should be relatively straightforward to assign PIDs to a majority of the RI data objects. A profound gap in knowledge about what persistent and unique identifiers are, what they can be used for, and best practices regarding their use, emerged. Most identifier systems used are based on handles (DOIs from DataCite were the most common, followed by ePIC PIDs), but some RIs rely on formalised file names. While a majority see a strong need for assigning PIDs to their "finalised" data (individual files and/or databases), few apply this to raw data, and even fewer to intermediate data – indicating that PIDs are not used in workflow administration. Also, metadata objects are seldom assigned PIDs. Costs for maintaining PIDs are typically not treated explicitly by RIs.

3.1.1 Data citation requirements

Currently, users refer to datasets in publications using DOIs if available, and otherwise provide information about producer, year, report number etc. either in the article text or in the References section. A majority of the RIs feel it is absolutely necessary to allow unambiguous references to be made to specified subsets of datasets, preferably in the citation, while few find the ability to create, identify, and later cite collections of individual datasets is important. Ensuring that credit for producing (and to a lesser extent curating) scientific datasets is "properly assigned" is a common theme for all RIs – not the least because funding agencies and other stakeholders require such performance indicators, but also because individual PIs want and need recognition for their work. Connected to this, most RIs have strategies for collecting usage statistics for their data products, i.e., through bibliometrics searches (quasi-automated or manual) from scientific literature, but thus often rely on publishers indexing also data object DOIs. NOTE: RIs were asked to characterise their "designated user community" needs, but most responded with RI-centric requirements. This may be because there was not sufficient opportunity to directly communicate with users. Normally, the RIs' highest priority is to improve

their productivity, in this case by having as much of their data identification and citation automated.

3.1.2 Assessment of the I&C requirements

The *Identification and Citation* requirements that are summarised in the previous section validate the need for this provision in ENVRIplus. However, the RIs showed significant diversity in their data-identification and data-citation practices and many were not aware of their importance in supporting data use. *Data Identification and Citation* are, however, key to reproducibility and quality in data-driven science and very often vital in persuading data creators of the value of contributing their data, data users of the need to recognise that contribution and funders to continue to support data gathering and curation.

Many researchers today access and therefore consider citing individual files. This poses problems if the identified files may be changed, the issue of *fixity*. Many research results and outputs depend on very large numbers of files and simply enumerating them does not yield a comprehensible citation. Many derivatives depend on (computationally) selected parts of the input file(s). Many accesses to data are via time varying collections, e.g., catalogues or services, that may yield different results or contents on different occasions — generically referred to as *databases*. Some results will deal with continuous streaming data. Often citations should couple together the data sources, the queries that selected the data, the times at which those queries were applied, the workflows that processed these inputs and parameters or steering actions provided by the users (often during the application of the scientific method) that potentially influenced the result. All of these pose more sophisticated demands on the *Data Identification and Citation* systems. In due course, those advanced aspects that would prove useful to one or more of the RI communities should be further analysed and supported. This is revisited in the technology review in **Chapter 5.1** below.

In summary, the use of persistent and unique identifiers for both data and metadata objects throughout the entire data life cycle needs to be encouraged, e.g., by providing training and best-use cases that illustrate both what should be done, and how to do it with what methods and applications. Indeed, the RIs should become apt users of both tools for applying PIDs to their (digital) resources and data, as well as tools that use supported identification & citation mechanisms to facilitate research work — including data movement, method application to sets, provenance tracing, etc. [Myers 2015]. There is strong support for promoting “credit” to data collectors, through standards of data citation supporting adding specific sub-setting information to a basic (DOI-based) reference. Demonstrating that this can be done easily and effectively, and that data providers can trust that such citations will be made, should be a priority, as it will lead to adoption and improvement of FAIR (and fair!) citation practices.

4 COMPONENTS OF PID IMPLEMENTATION, DATA PUBLISHING AND CITATION

In this Chapter, we cover some of the most important components of PID implementation, data publishing and data citation and usage statistics that need to be considered by research infrastructures. The list is in no way complete, but can be seen as a starting point for the RIs’ own information gathering and planning processes on the way to implement the system design and best practices for identification and citation that are presented below in **Chapter 8**.

4.1 The Persistent Identifier zoo

In this section, we present an overview of seven of the most commonly used persistent identifier types. The underlying study was performed in the summer of 2016 by Huber and co-workers, and the numbers and statistics represent the status of the re3data.org registry (http://www.re3data.org/) at that time.

4.1.1 The Handle System (HS)

Arguably the biggest impact in the field of persistent identification of digital research resources was achieved by the Handle system [Kahn and Wilensky 1995]. The Handle System (HS) describes a minimal set of requirements for an infrastructure for the identification of objects in a digital infrastructure and how the identity of an object can be related to its location. The system is agnostic to the contents of the objects, keeping it open for interoperability with future applications. The Handle system separates the identifier from the resolving mechanism, making it independent of HTTP and DNS but in practice, the system is mostly leveraged using a HTTP proxy that allows the use of a RESTful API and URLified handles. The Handle system supplies a stable, distributed platform for the resolution of identifiers to URLs, including methods more sophisticated than HTTP redirects like template handles and embedded metadata.

In the sample of 1381 repositories listed in the re3data repository at the time of the study, the Handle system is used by 102 repositories. Handle is mainly used by institutional repositories, which might be linked to the role of Handle as an identifier in repository software like DSpace (see http://www.dspace.org/).

Besides the governance of top-level namespaces the HS does not provide more than the technical platform and comes with no obligations with respect to policies, for instance towards the persistence of the resolution of identifiers towards their targets.

4.1.2 Digital Object Identifier (DOI)

Looking at the 475 repositories using any kind of PID system, the most commonly implemented identifier type was the digital object identifier (DOI). DOIs, which were introduced in 1998 by the International DOI Foundation (see http://www.doi.org/), were used by 275 out of those 475 repositories, meaning that the use of the DOI eclipsed all other persistent identifiers. The use of DOI persistent identification of data initiated by a project funded by the German research foundation in 2003 [Klump et al., 2016]. DOI were chosen because of their already established part in the scholarly publication infrastructure.

The Digital Object Identifier (DOI) makes use of the Handle system and uses its namespace “10.[subnamespace]/”. DOI distinguished from other uses of the handle system by the underlying social contract. In this social contract participating parties pledge to maintain the resolution of identifiers to web endpoints indefinitely. This means that identifiers will theoretically always resolve to somewhere even though the referenced object might no longer exist. (See **Chapter 4.3** and **Chapter 4.4** for a discussion of “tombstones”.)

4.1.3 Uniform Resource Name (URN)

The origin of the Uniform Resource Name (URN) is its historical use as a name for a Uniform Resource Identifier (URI) that uses the URN scheme. Defined in 1997 in RFC 2141 [Moats 1997], URNs were intended to serve as persistent, location-independent identifiers, allowing the simple mapping of namespaces into a single URN namespace. The existence of such a URI does not imply availability of the identified resource, but such URIs are required to remain globally unique and persistent, even when the resource ceases to exist or becomes unavailable.

Primary advocates for the use of URN were the national libraries. Unlike the handle system, URNs do not use a common resolver system. It is therefore up to the user to know which resolver system to use. This has been a severe impediment against the uptake of this system.

URN systems were also primarily offered by national libraries, which offered great organisational stability to the system. In working with data centres the national libraries showed difficulties in adjusting their business processes to the requirements of data centres. This resulted in an overall low uptake of URN as a persistent identifier for data and only 16 out of 475 repositories make use this identifier system.

4.1.4 Persistent URL (PURL)

Persistent URL (PURL) was intended by the Online Computer Library Centre, Inc. (OCLC) as a bridging technology to prepare for the introduction of Universal Resource Names (URN). PURL implements the URI concept. It does not separate between identifier and resolving mechanism. PURL has no single global resolving mechanism and PURL resolvers do not communicate amongst each other and share resolving information like DNS servers do. PURL has little social infrastructure and formal governance. In 2014 OCLC withdrew its institutional support and the future of PURL was unclear. For some time PURL experienced severe technical problems. As a consequence, the system was put into a 'read-only' maintenance mode[1]. In September 2016, however, it was announced by the OCLC and the Internet Archive that the URL redirection service, on which PURL is based, will in future be operated by the Internet Archive[2]. This move brought PURL back from the brink of extinction. In December 2015 a total of 16 research data repositories in re3data.org were listed as using PURL, and only few of them using PURL exclusively.

4.1.5 Archival Resource Key (ARK)

The ARK (Archival Resource Key) identifier system has been introduced by the California Digital Library in 2001 [Kunze 2003]. ARKs are currently used by eleven repositories listed in the re3data registry. It was developed as an alternative to schemes like PURLs, URNs and Handles. The founders argued that persistence of identifiers is a matter of service and reliant on the continued stability and support of the service behind the identifiers. Initially the system was planned to enable decentralised resolvers, however this principle was never realised. Instead, ARK resolution depends on local resolvers that ARK-issuing archives have to provide and maintain, and a central service called N2T (name to things) which however only resolves ARKs which are registered with the University of California. Thus ARKs rather represent a resolver convention and an associated syntax for the ARK URI pattern.

4.1.6 Life Science Identifier (LSID)

Life Science Identifiers (LSIDs) have been introduced by the Object Management Group (OMG) in 2004. Since 2009 the biodiversity informatics communities' standardization authority (Taxonomic Database Working Group, TDWG) strongly supports LSIDs as the preferred GUID (Globally Unique Identifier) technology. LSIDs are now used by all globally leading providers for biodiversity data to identify organism names. LSIDs do not provide a global resolving mechanism nor centralised provider registration. The implementation of this standard is relatively complex, as resolution is DNS based and requires a multistep procedure. Furthermore, the associated metadata format is RDF (Resource Description Framework) triples. Consequently, the technology is controversially disputed (see, e.g., [Hyam 2015] or [Page 2015]) and opinions in social media tend to favour a simpler identifier system such as HTTP URIs. In 2016, maintenance on TDWGs LSID resolution service was ceased and TDWGs support of LSIDs was heavily questioned by members of the group. After two months without a central resolving system, a resolver was made available at http://www.lsid.info. However, the discussion is continuing and, as parts of the biodiversity informatics community still feel uncomfortable with LSIDs, they are recommending a switch from LSID to CoolURI [Guralnick et al., 2015]. This approach may however be fragile, because it assumes that the base URL will not change, since this is a prerequisite for the operation of the system.

4.1.7 "Cool" Uniform Resource Identifiers (CoolURIs)

Compared to the strict criteria of Nestor [Bütikofer 2009] and other related efforts, "cool" (meaning unchanging or static) Uniform Resource Identifiers (CoolURIs) somehow represent an

[1] See announcement made via the now archived JISCmail discussion message https://www.jiscmail.ac.uk/cgi-bin/webadmin?A2=ind1511&L=DC-ARCHITECTURE&F=&S=&P=3711.
[2] See https://www.oclc.org/news/releases/2016/201623dublin.en.html.

anarchic view on identifiers. Similar to URN, the idea of CoolURIs goes back to early ideas about identification and location of objects on the web. The idea of CoolURI [Berners-Lee 1998] is fundamental for the Semantic Web. It is based on Uniform Resource Identifiers (URI) which, by proclamation, will not change. They make use of standard HTTP functionalities, in particular content negotiation [Wikipedia 2017a], to enable the URI to be resolved to different representations (RDF, HTML) of the same object. CoolURIs allow webmasters to maintain the persistence of their resource identifiers, the URIs, with a minimum of effort and without a centralised PID system.

Advocates of the CoolURI system reasoned that the use of HTTP functionalities is a bonus, suggesting that URI should be actionable. However, over the years this has proven to be unstable, the main reason for this being the fragility of base URL. The result of unstable base URLs will be "link-rot on steroids". There is already anecdotal evidence of base URL failures from the validation of xml schemas in long-term archiving of XML documents by the national libraries.

The CoolURI concept relies on HTTP as resolving mechanism and assumes that the HTTP protocol will be around for a long time. HTTP went into operation in 1991 and on the timescales of technical developments in information technology we should not assume that it will still be in use in 25 years' time.

4.2 Identifiers for non-data entities

Persistent identifiers are useful for many other entities than data objects and scientific articles. In the following, we list a selection of such entities which have a special interest to ENVRIplus partner RIs.

A. Identifiers for people

During the last five years, more and more researchers have become used to registering with ORCID and then using their ORCID IDs for communications with journal publishers, their funding agencies and in other research contexts. However, also other individuals associated with research projects (and active in producing research outputs) – such as research engineers, data curators, programmers and many others – should also be encouraged to sign up for ORCID or similar persistent identifiers schemes for individuals like ISNI. The personal IDs can then be stored in RI catalogues, and be included in metadata objects and DataCite records. For more information, see **Chapter 6.1.3** (ORCID) and **Chapter 6.1.4** (ISNI).

B. Identifiers for organizations

Also the organisational entities involved in research projects should in principle obtain persistent identifiers, for example via ISNI. However, this may not be as simple and clear-cut as for persons, since reorganisations and restructuring may occur at any time. For more information about ISNI, see **Chapter 6.1.4**.

C. Identifiers for instrumentation and sensors

By assigning unique and persistent identifiers to sensors and other instrumentation, and using these PIDs consistently in both cataloguing and curation, researchers can simplify the management and collection of observation metadata records, and facilitate property lookup and provenance tracing throughout all steps of the research data processing cycle. See **Chapter** Error! Reference source not found. (French marine research cruises) and **Chapter** Error! Reference source not found. (best practices) for more details.

D. Identifiers for physical samples

In order to simplify the referencing of physical samples, they can be registered and assigned a unique identifier. One initiative that provides this possibility in Earth sciences is System for Earth Sample Registration (SESAR), which allocates IGSNs (International Geological Sample Numbers) to environmental samples. (See http://www.geosamples.org/igsnabout for more details.)

E. Identifiers for data content types

In order to facilitate (re-)use of datasets, especially in the context of machine-actionable workflows, it is useful to make use of persistently identified Data Type definitions. These should include a basic description of the characteristics of a given data or variable, but can also contain information on which software should be used to process it further. See e.g. the recommendations of the RDA Data Type Registries working group (https://www.rd-alliance.org/group/data-type-registries-wg/outcomes/data-type-registries) and also **Chapter** Error! Reference source not found. below.

F. Identifiers for software

GitHub (https://github.com/) and similar software repositories support versioning, and as such allow the code author to link directly by URL to a specific code package or file. In GitHub, objects can themselves be linked to dataset DOIs, so there are possibilities of cross-referencing. However, at the moment it is not yet possible to provide a DOI or any other PID to software codes or packages in GitHub. Notably, the German Climate Computing Centre DKRZ (see **Section 7.1.2**) is about to apply for national project funding to offer sustainable production and long term storage of scientific software. This will account for versioning and include the use of persistent identifiers. See also **Chapter 8.3.1**.

G. Identifiers for workflows

Workflows and workflow engines are being increasingly used also in environmental and Earth sciences as a means of organising and sharing scientific computations and analysis procedures. Referring to specific workflows simplifies the collection of provenance records associated with datasets. Registering workflows and assigning them PIDs promotes efficient documentation of workflows, allows making unambiguous references to them in e.g. provenance descriptions, and supports their reuse by both humans and machines. See **Chapter 8.3.1** for more information.

4.3 Landing pages

[Starr 2015] provides a summary of best practices surrounding the use and application of landing pages. In order to provide an optimally interoperable environment, in which both humans and machines can efficiently interpret and act on the information about a DO returned by a PID resolving agent, the persistent identifier used in a (data) citation should not point directly to the digital object itself, but to a so-called landing page (or set of pages). In effect, the landing page should be a set of information — typically presented as a web page that can be either static or created on-the-fly from a cataloguing service — that provides information about the data object including structured metadata and unstructured text and other content.

This landing page should persist, even when the digital object no longer exists or is inaccessible, and continue to provide at least basic information about the data — including the reasons for its removal, or links to any succeeding version. This is in line with the JDDCP, which proposes that the metadata should be curated in a way that ensures they remain a part of the citable scholarly record for a foreseeable future.

The landing page also serves as an “access gateway” for data objects that may not be accessible to every user — e.g. because of licensing, confidentiality or other reasons that require a user to pass an identification/authentication/authorization filter. Those parts of the associated metadata that are openly accessible, or authorised for the current user, can then be displayed on the landing page, together with information about the restrictions. Thirdly, by use of content negotiation [Wikipedia 2017a], different encodings of a data object (as well as its related metadata) can be served, depending on the settings of the user agent that visits the landing page.

[Starr 2015] recommend that the following information is provided on the landing page:

Recommended

- Dataset description
    - Dataset identifier
    - Title
    - Description
    - Creator (ORCID if available)
    - Publisher/Contact
    - Publication date/Release date
    - Version
- Persistence statement
- Licensing information

Conditional:

- Version
- Access controls
- Data availability and disposition

Optional

- Explanatory or contextual information
- Links to tools or other software

4.4 "Session PIDs" vs "Citation PIDs"

Is it then really useful and meaningful to define large numbers of *persistent* identifier records to datasets that one knows *a priori* will not be existing for a very long time, but perhaps be deleted after a few months? If it is, does it matter which of the different available types of persistent identifiers (see Chapter 4.1 below) one chooses for these "short shelf-life" data? And who should pay for maintaining landing pages for obsolete datasets (also called "tombstone pages")?

Connected to the discussion around these questions, the terms "session PID" and "citation PID" have emerged [Bilder 2016: Geoff Bilder's presentation "Layered services on top of PIDs", RDA Europe training course & workshop "Views about PID systems", Garching, Germany, Aug. 31-Sept. 2, 2016]. Here, "session PID" refers to persistent identifiers that are applied to data objects that need to be referenceable, for example raw sensor data, or intermediate data produced during evaluation and analysis processes, that should be easily and unambiguously referred to in the provenance record of a finalised data product. Here, for example a PID from ePIC (see Chapter 6.1.1) would be quite sufficient. Conversely, a "citation PID" should be used for data that are "publishable" and ready to be used for research purposes, and therefore may warrant the assignment of an identifier that is associated with a richer metadata schema, such as a DataCite DOI.

It should also be remembered that in a strict sense, the word "persistent" in the term "Persistent and Unique Identifier" refers to the identifier, and not to the object that is being identified. There is a common misconception that once a (data) item has been assigned a PID, the item may not be deleted from the location that the PID's metadata resource locator field points to. This is not correct, as discussed by, e.g., [Duerr 2011], who point out that since there are costs associated with maintaining datasets, it may be deemed desirable or even necessary to delete obsolete, or otherwise replaced and updated, objects. (The maintenance costs for long-term storage at a high-quality, secure and trusted repository can be high, and may be difficult to cover for more than a few years after the termination of a research project.) However, it remains important to ensure that the identifier(s) that were assigned to the old data should be maintained, especially if the data were used, or referred to, in published works. To ensure that the identifiers remain resolvable, the location pointer in the PID registry should be updated with a link to a "tombstone page", i.e. a (preferably machine-actionable) landing page which indicates

that the dataset has been deleted, and redirects (if applicable) to any updated version. If possible, the tombstone landing page should also provide metadata including the reasons for deletion, as well as provenance and context information for the missing data [Duerr 2011].

4.5 Data publishing

The European Commission and many other organisations propose the practice of open science, which can be described as including open access, open data, open reproducible research, open science evaluation, open science policies, and open science tools [FOSTER 2016]. Similarly to the open access (for publications) and open software (for code) concepts, open data typically means that data are open to use with minimal restrictions – but it does not have to be free of charge, i.e. gratis, to quote [Borgman 2007]. A major difference between publishing data and a research publication is that data are not static in the way that a published paper generally is – data can be dynamic, "living", and evolving by new versions being made available [Austin 2016]. Publishing data could mean posting a dataset on a website, but often involves far more work than that, such as making considerations regarding data formats, metadata descriptions, and persistence of the data [Callaghan 2013].

The benefits of publishing data include allowing reproducibility of research, increasing the quality of research, creating new knowledge by combining datasets in new ways, and to make data accessible and usable to a wider audience, which could lead to more scientific publications [Austin 2016], [Borgman 2007], [Borgman 2015]. The audience of published data varies depending on topic and other characteristics of the research and the data, and could include researchers in the same field, researchers in related or unrelated fields, policy makers, industry, students and educators, or the general public. There are possibilities of setting access levels to published data, depending on the publishing venue chosen. Whereas openly accessible data might be the most common approach, some repositories allow for registering metadata that are openly accessible, while the data are either embargoed (meaning, accessible after a chosen time period), or available upon request. This relates to licensing of data, which, if a set of conditions are met, can grant permission to users (other than the rights owner) to utilise data.

The most widely used licenses for open data today are the ones from Creative Commons (https://creativecommons.org). Licenses typically contain requirements regarding Attribution (others are required to give credit to the rights owner as determined by the license), ShareAlike (whether others are required to distribute a derivative work under the same license), NonCommercial (whether others are required to only use this work non-commercially), and NoDerivatives (whether modifications are allowed, or only copying, distributing and performing the original version of the work) [Creative Commons 2016a]. The most permissive form is CC0, which means that the rights owner waives all rights to the data (used by Europeana – Europe's digital library, figshare [Creative Commons 2016b], and the ATLAS community at CERN [Halperin 2016]) but still disclaims responsibilities for it. The second most permissive Creative Commons license is CC-BY which allows others to share and adapt the data under the condition that the owner of the right is attributed (which could be a citation for data) [Creative Commons 2016c]. Adding the suffix –SA, for Share Alike, to the CC-BY license adds the condition that the data are distributed under the same license as the original [Creative Commons 2016d]. By adding the prefixes of choice, it is possible to grant the permissions that suit the data or any other works of relevance.

Austin and colleagues [Austin 2016] have proposed a set of key components for data publishing (see **Figure 5**) based on their analysis of workflows in 25 repositories, data related projects and data publishing platforms, all within the work of the RDA/WDS Publishing Data Workflows Working Group. The proposed required elements for a published dataset include a repository entry with a PID, metadata following a standardised schema, curation, and distribution. The optional elements have three themes: providing context, such as documentation and links to articles and other research objects; quality control, including peer review; and improving visibility and access to the data [Austin 2016].

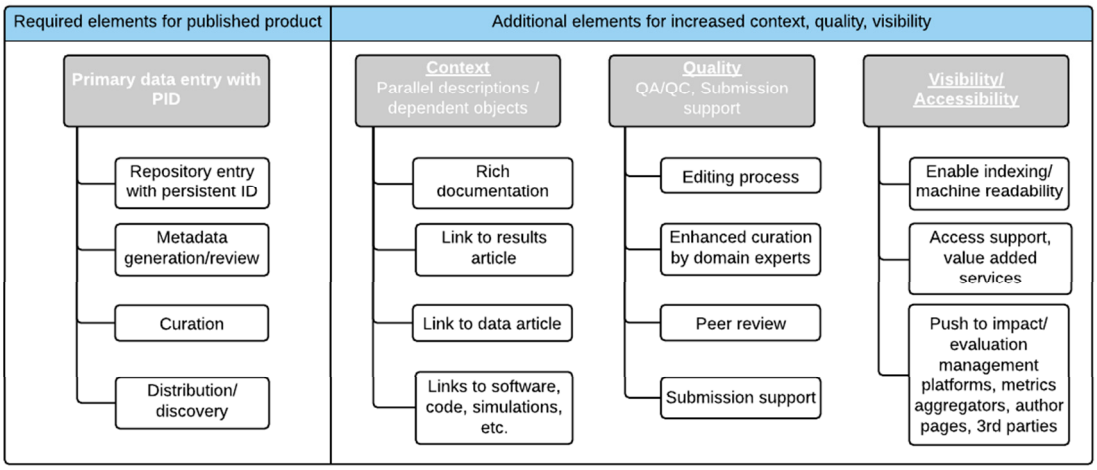

FIGURE 5. KEY COMPONENTS OF DATA PUBLISHING. REQUIRED ELEMENTS ARE SHOWN IN THE LEFT-HAND BOX, AND OPTIONAL COMPONENTS ARE IN THE RIGHT-HAND BOX. FROM [AUSTIN 2016]

The data workflow analysis [Austin 2016] resulted in the workflow description found in **Figure 6** below. The two most common types of data publishing were found to be submitting data to a repository, and submitting a data article to a data journal. Focussing on the submission of data, the workflow consists of ingestion into a repository, a process which includes creating a PID for the data. A review of the data is also part of the workflow, based on the policies of the repositories, and may include metadata, formatting and a host of other aspects of the data curation. To ensure the quality of data, it is recommended that quality assessment procedures are available to users. Quality control procedures for dynamic data are also identified as important, due to more data being shared earlier in the research process.

To help ensure that datasets are stored in an appropriate and sustainable manner (including the persistence of data over time), the authors recommend that repositories apply for certification from e.g. the Data Seal of Approval (DSA; see http://www.datasealofapproval.org/), and add that this could aid the data publishing process by facilitating collaboration with other actors involved in data publishing, including publishers. The repositories and other facilities analysed generally used a PID system, in most cases DOIs, and the authors note that the importance of data citation through linking PIDs was generally taken into account.

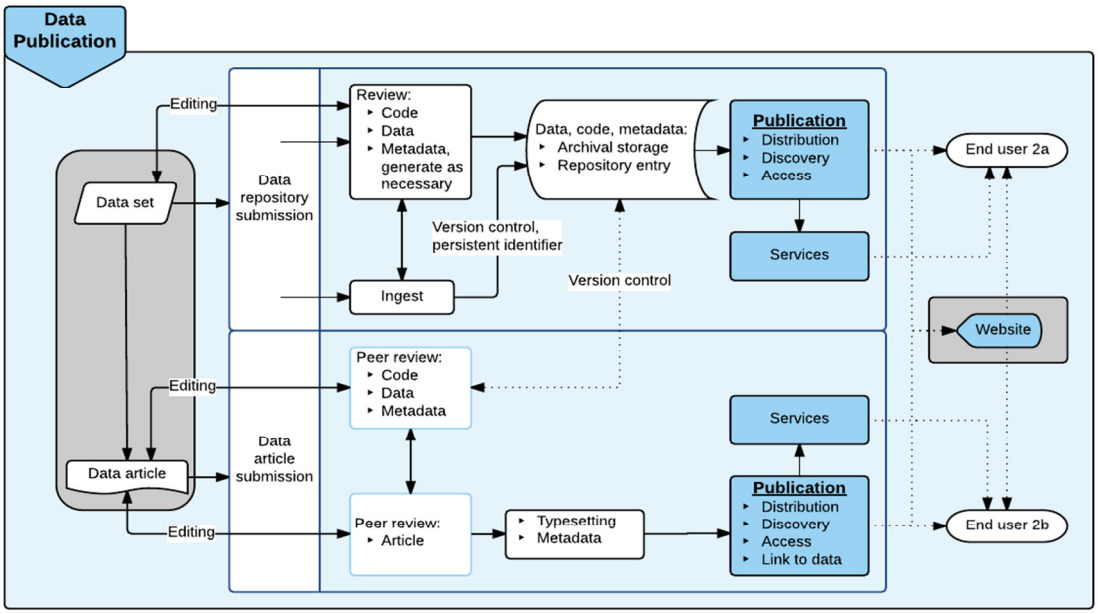

FIGURE 6. DATA PUBLISHING WORKFLOW [AUSTIN 2016]

Linking data with papers, people and other research objects through PID systems is a strong trend, which was shown by the many presentations at the conference on PIDs, PIDapalooza, 9-10 November 2016 in Reykjavik, Iceland. The conference found a common ground in the importance of PIDs for being able to identify, find, and link resources, and to reuse the metadata for published objects in other contexts. This could mean that the description of a project and the generated data at the planning stage, e.g. in a data management plan, can be extracted and used to generate content to a progress report to funders. Project THOR (see **Chapter 7.2.1**) is an example of an initiative to link data, publications and persons through their ORCIDs.

4.6 Citation analysis

Citation of research publications is one of the traditional ways of crediting authors and building on previous research [Cronin 2005]. In this context, a citation is a reference to another scholarly work, such as a journal article or monograph, and is found in-text and in the reference section of a scholarly publication. Citation analysis typically involves statistical methods for studying how scholars cite each other, the most commonly used methods being bibliometrics. The results of bibliometrics studies include author networks based on the frequency of citations between authors, as a tool for career advancement, and as a tool for distributing funding based on citation statistics on institutional levels. The latter type of analysis can also be labelled as scientometrics, a field which focuses on measuring science, including journal impact factor which is a commonly used metric by scholars for choosing an appropriate publishing venue.

A higher journal impact factor is seen to reflect a higher quality journal, which could aid career advancement, funding opportunities and status in academia. Journal impact factor is increasingly criticised for focussing research assessment on journal metrics instead of the quality of the research itself. This is reflected in the San Francisco Declaration on Research Assessment, DORA [DORA 2012], which is a call for using other ways of assessing the quality of research. The recommendations of the DORA are aimed at different stakeholders in academia, including the practices of researchers, funding agencies, and publishers, and three themes are found throughout the recommendations: promoting a broader spectrum of metrics to move from journal-based metrics as predominant, making research assessment on factors other than journal metrics, and making use of the opportunities presented by digital publications, e.g. removing limits on the number of references allowed in journal articles [DORA 2012]. The altmetrics concept offer alternative metrics for scientometrics and bibliometrics, see Chapter 6.3.4, as does the Making Data Count project (http://mdc.lagotto.io/) focussing on metrics for data. Citing of research objects other than publications is becoming more prevalent, data perhaps getting the most attention due to the open science and open data trends throughout the world in the latest years.

5 TECHNOLOGIES FOR IDENTIFICATION & CITATION

This Chapter is based on (but not identical to) Identification & Citation-relevant sections of the ENVRIplus Work Package 5 deliverable D5.1 [Atkinson 2016]. The material is included here in order to provide context and background to both **Chapter 8.3** and **Chapter 9.3**.

5.1 Technology review

It is important to keep in mind that there are many different actors involved in *data identification and citation*: data producers (RIs, agencies, individuals); data centres (community repositories, university libraries, global or regional data centres); publishers (specialised on data, or with a traditional focus); and data users (diverse ecosystem, from scientists, experts to stakeholders and members of the public). The deployed technologies should reflect the needs and requirements for all of these. Here the focus is on RIs that typically involve all of those viewpoints. Time constants for changing old practices and habits can be very long, especially if they are embedded in established cultures or when capital investment is required.

For these reasons, updating, or implementing totally new, technology alone does not improve "*usage performance*"[3], as the adoption behaviour of the "*designated scientific community*" will influence the discoverability and ease of reuse of research data. Scientific traditions and previous investments into soft- or hardware can lead to large time constants for change. Adopting new database technology quickly could, on paper, provide large benefits (to the data providers) like lower costs and easier administration and curation, but may *de facto* be unacceptably lowering overall productivity for significant parts of their user community over a long period of time while the transition is achieved.

5.1.1 Two-to-five year analysis of state of the art and trends

As evident from the large number of on-going initiatives for applying identifiers to, and subsequently providing linkages between, all components of research – from individual observation values to the people making them – it is a very difficult task to even try to envisage how the data-intensive research landscape will look in a few years from now.

Here, we list some of the issues and ideas that are being worked on now, and which we -- based on discussions with ICT experts form Work Package 5 and the outcomes of the Requirements survey presented in **Chapter 3** -- feel will continue to be of importance in the coming years:

A. A majority of (starting-up) RIs adapt data curation strategies that are fully capable of handling dynamic data (both versioned static files and truly dynamic databases), centred around persistent identifiers for both data & metadata objects and queries.
B. Standards for unambiguous referencing of subsets of datasets (in citations and in workflow contexts) will become widely adopted by scientists and publishers alike, enabling both efficient (human and machine) extraction of "slices" of data as well as detailed (micro)attribution of the producers of the data subset.
C. More complex data objects will become common, including data collections, "research objects" containing both data and related metadata, and other (virtual) aggregates of research information from a multitude of sources. This will require new strategies for content management and identification at both producer and user level.
D. Systems for allocating persistent identifiers will become more user-friendly, e.g., by development of human-oriented user interfaces (UIs) to complement the APIs (Application Programming Interfaces) that are common to all major identifier registries. This will have profound positive impacts on the administration and reproducibility of scientific workflows.
E. To enable efficient automation of data discovery and processing, it will become common to store an enhanced set of metadata about the objects directly in the PID registries' data bases, e.g., related to fixity, versioning, basic provenance and citation.
F. The current trend to implement an ever tighter automated information exchange between publishers, data repositories and data producers will continue, and become the norm in many fields including Environmental and Earth Sciences.
G. More effective usage tracking and analysis systems that harvest citation information not only from academic literature but from a wide range of sources will be developed.

Individual ENVRIplus RIs are engaged in a number of the above-mentioned developments through the activities outlined in the Description of Work of several work packages in Themes 1 and 2.

There is also active participation, by individual ENVRIplus RIs, in projects such as EUDAT2020 or as use cases in RDA groups, see **Chapter 6.4.1**. However, the relatively short lifetimes, and limited number of members, of this type of project or group often has several negative consequences. Firstly, there may not be enough diversity within the use cases to encourage the development of broad solutions that cover the needs and requirements of a wider range of

[3] The working practices actually adopted by the practitioners in all of the roles involved with data or the work that created it or that it is used for.

communities. Secondly, the knowledge and experience gained through such work often ends up benefiting only a small number of RIs – if there is any long-lasting application at all!

ENVRIplus could therefore make a difference by setting up a (project-independent) platform for informing practitioners about on-going initiatives (especially those that involve ENVRIplus members, but not as part of ENVRIplus itself), collection of RI use cases for passing on to the technology developers, and finally promoting the dissemination, implementation and uptake of effective examples. It has recently been proposed to set up such a group (the Information Systems Strategy and Engineering Group, ISSEnG) that would offer ENVRIplus partners a platform to discuss data and technology issues with a longer-term horizon.

5.1.2 Details underpinning the two-to-five year analysis

In this section, we present more background for the 7 topics (A-G) listed above. For each topic, some specific examples of relevant technologies are listed, together with a brief narrative discussion and suggestions for further reading.

A. A majority of (starting-up) RIs adapt data curation strategies that are fully capable of handling dynamic data (both versioned static files and truly dynamic databases), centred around persistent identifiers for both data & metadata objects and queries.

- Main technology needs: versionable databases to support "time machine" retrieval of large datasets (also sensor data) that are dynamic.

Sources: [Rauber 2016] and personal communications with A. Asmi, 2016.

There exist already today several database solutions that support versioning of database records—both SQL and NoSQL-based. Both approaches have advantages and disadvantages, but with optimised and well-planned schemas for storing all transactions and their associated timestamps, it is possible to achieve "time machine"-like extraction of data (and metadata) as they existed at any given time, without significant losses in performance – at least for moderately-sized databases. But challenges remain, e.g., for databases required to store long time series of high-frequency sensor data. For data stored as flat files, it is mainly the metadata that must be stored in a database supporting versioning, to allow identification of what file(s) represent the "current state" of the data at a given point in time. (There remains, however, the issue of what a "point in time" actually means, as it may refer to the data collection or processing, the data publishing, any update to the data etc.)

- Connections to cataloguing and maintenance of provenance records, supporting automated metadata extraction and production for machine-actionable workflows.

Sources: [Tilmes 2010], [Duerr 2011] (see example in the article supplement!)] as well as on-going work in RDA Metadata Interest Group, RDA Research Data Provenance Interest Group and EUDAT2020.

In order for data-driven research to be reproducible, it is an absolute requirement that not only all analysis steps be described in detail, including the software and algorithms used, but that the input data that were processed are unambiguously defined. Ideally, this is achieved by minting a persistent identifier for the dataset as the basis for the citation, and then adding details about the date when the data was extracted, the exact parameters of the subset selection (if used), version number (if applicable) and some kind of fixity information, like a checksum or content fingerprint. Optimally, at least one of 1) the citation itself; 2) the PID record metadata and/or 3) the resource locator associated with the PID, will provide all this information in a machine-actionable format, thus allowing workflow engines to check the validity and applicability of the data of interest.

Currently, a majority of the ENVRIplus RIs – and their intended user communities – haven't yet started to implement the outlined practices in a consistent manner. As a consequence, the reproducibility of research based on data from these RIs could be called into question. What is needed to change this situation, are good examples and demonstrators that can be easily

adopted by RIs (without much investment in time and software). Such best practices need to be developed in cooperation across the Work Packages of Theme 2.

B. Standards for unambiguous referencing of subsets of datasets (in citations and in workflow contexts) will become widely adopted by scientists and publishers alike, allowing both efficient (human and machine) extraction of "slices" of data as well as detailed (micro)attribution of the producers of the data subset.

- Query-centric citations for data, allowing for both unambiguous and less storage resource-intensive handling of dynamic datasets

Sources: [Duerr 2011], [Huber 2013], [Rauber 2016]

Datasets from research may undergo changes in time, e.g., as a result of improvements in algorithms driving a re-processing of observational data, errors having been discovered necessitating a new analysis, or because the datasets are open-ended and thus being updated as new values become available. Unless great care is taken, this dynamic aspect of datasets can cause problems with reproducibility of studies undertaken based on the state of the dataset at a given point in time. The RDA working group on Data Citation has therefore produced a set of recommendations (in 14 steps) for implementing a query-based method that provides persistently identifiable links to (subsets of) dynamic datasets. The WG has presented a few examples of how these recommendations can be implemented in practice, but there is a great need for continued work towards sustainable and practical solutions that can easily be adopted by RIs with different types of data storage systems.

C. More complex data objects will become common, including data collections, "research objects" containing both data and related metadata, and other (virtual) aggregates of research information from a multitude of sources. This will require new strategies for content management and identification at both producer and user level.

- systems for cataloguing and handling more complex collections, both of datasets and metadata (c.f. "research objects").

Sources: OKFN, wf4Ever, the RDA Data Collections WG (just starting) and RDA Data Type Registries WG (concluded with recommendations).

The increasing complexity of research data and metadata objects adds more challenges. Firstly, in contrast to printed scholarly records like articles or books, data objects are often in some sense "dynamic" – updates due to re-analysis or discovered errors, or new data are collected and should be appended. The content can also be very complex, with thousands of individual parameters stored in a single dataset. Furthermore, there is a growing trend to create collections of research-related items that have some common theme or characteristic.

In the simplest form, collections can consist of lists of individual data objects that belong together, such as 365 daily observations from a given year. Similarly, it may be desirable to combine data objects and their associated metadata into packages, or to create even more complex "research objects" that may also contain annotations, related articles and reports, etc. Collections can be defined by the original data producers, but may also be collated by the users of the data – and may thus contain information from a large variety of sources and types, i.e. forming virtual collections. This diversity is prompting work on providing tools for organising and managing collections, e.g., using APIs that are able to gather identity information about collection items (through their PIDs), as well as minting new PIDs for the collections themselves.

Additionally, there are not yet rules established how to generate metadata of collections. In the world of printed media the creator of a collection sometimes may be referred to as editor, collector, or may be the director of a library. In the world of data, however, creators mostly do not yet have roles: not in metadata schemas, nor in normalised transportation xml interfaces.

Furthermore, there is also a need for sustainable registries for data type definitions that can be applied to "tag" content in a way that is useful and accessible both to humans and for machine-

actionable workflows. However, the use of data types varies greatly between different user communities, making it a difficult task to coordinate both the registration of definitions as well as a sustainable operation of the required registries, especially if these are set up and operated by RIs. Here more work is needed in collaboration with a number of RIs each with differing data-set structures and catalogue organisations, in order to provide clear recipes for data typing.

D. D. Systems for allocating persistent identifiers will become more user-friendly, e.g., by development of human-oriented user interfaces (UIs) to complement the APIs (Application Programming Interfaces) that are common to all major identifier registries. This will have profound positive impacts on the administration and reproducibility of scientific workflows.

- Adoption of a common API for PID minting, applicable across registries and methods.

Sources: [Duerr 2011], [Socha 2012], [Klump 2015] as well as work by the RDA PID Information Types WG (concluded) and the RDA PID Interest Group (in progress).

Although a number of systems for persistent identification of e.g., scientific publications have been available for over a decade, relatively few researchers are consistently applying these systems to their research data. There is, at the same time, a pressing need to encourage data producers to mint PIDs for any (digital) items belonging in the research data lifecycle that should be "referable" – including also raw data and datasets produced during analysis, and not just finalised and "published" datasets. Surveys have indicated that the reasons for the slow adoption rate include a lack of knowledge about the existing opportunities, confusion over their relative differences and merits, and difficulties related to the identifier minting process (especially when it needs to be performed on a large scale, as is often the case for data). The latter problem is to a large extent due to the large variety in design and functionality of PID registry user interfaces and APIs, and there are now several initiatives looking into how the registration and maintenance of PID records can be streamlined and simplified. However, the proposed inclusive user and programmatic interfaces will need extensive testing by a wide range of different user communities. There are also institutional issues, including concerns over intellectual property rights that may inhibit the adoption of working practices or the delegation of authority to allocate PIDs.

E. To enable efficient automation of data discovery and processing, it will become common to store an enhanced set of metadata about the objects directly in the PID registries' databases, e.g., related to fixity, versioning, basic provenance and citation.

- Handle registries also need to become federated, and allow users to add community- or project-specific metadata to the Handle records (see recommendations of the RDA WG on PID information types), including those required for identity and fixity verification.

Sources: RDA PID Information Types WG (final), new RDA Data Collections WG (in progress) and presentations from the ePIC & DataCite PID workshop in Paris, 2015[4].

Mainly motivated by a desire to speed up and facilitate the automation of data discovery and processing, there are calls for the centralised Handle (and other PID system) registries to also allow data producers and curators to store more types of metadata about the objects directly in the registries' databases. Examples include information related to data content type(s), fixity, versioning, basic provenance and citation. This would speed up data processing since the requesting agent (e.g., a workflow process) would be able to collect all basic metadata via just one call to the PID registry, instead of needing to first call the registry and then follow the resource locator pointer to e.g., a landing page (from which data would need to be harvested and interpreted).

[4] See http://blog.datacite.org/recap

Some PID management organisations, such as DataCite (and the DOI Foundation) already support a relatively broad range of metadata fields, but other registries are more restrictive. The technology for storing the metadata is already in place, but database systems would need to be upgraded to allow for more PID information types. Also, registry servers' capacity to handle the expected large increase in lookup query requests must be upgraded. Optimal performance will require the PID information types themselves to be defined and registered in a persistent way, e.g., using a data type registry.

F. The current trend to implement an ever tighter automated information exchange between publishers, data repositories and data producers will continue, and become the norm in many fields including Environmental and Earth Sciences.

- Expanding the application of persistent unique identifiers for people and institutions in research data object management, including metadata and PID registry records.

Sources: ORCID and DataCite, THOR website and webinar series.

Driven by demands from large scientific communities (e.g., biochemistry, biomedicine and high energy physics), publishers and funding agencies, there is a strong movement towards labelling "everything" and "everyone" with PIDs to allow unambiguous (and exhaustive) linking between entities. Currently it is quite common for individual researchers to register e.g., an ORCID identity, and subsequently use this to link to articles in their academic publications record. This could be equally well applied to (published) research data, for example by entering ORCID IDs in the relevant "author" metadata fields of the DataCite DOI registry record, and allowing this information to be harvested by Crossref or similar services.

Connected with this is a growing trend to implement tighter information exchange (primarily links to content) between publishers, data repositories and data producers. There are several on-going initiatives looking into how to optimise and automate this, including OpenAIRE and the THOR project (coordinated by the British Library), which involves amongst others ORCID, DataCite and PANGAEA. It is expected that the outcomes of these efforts will set the norm.

However, to be fully inclusive and consistent (from a data curation and cataloguing point of view), this practice should be extended to all relevant "personnel categories" involved in the research data life cycle, including technicians collecting data, data processing staff, curators, etc. – not just principal investigators and researchers. This would allow both a complete record of activities for individuals (suitable for inclusion in a CV), but conversely can also be seen as an important source of provenance information for linked datasets.

G. More effective usage tracking and analysis systems that harvest citation information not only from academic literature but from a wide range of sources will be developed.

- Discovering and accounting for (micro)attribution of credit to data producers and others involved in the processing & management of data objects – especially in the context of "complex" data objects

Sources: [Uhlir 2012], [Socha 2012], [Huber 2013] and RDA Research Data Collections Interest Group

There is strong encouragement from policy makers and funding agencies for researchers to share their data, preferably under Open Access policies, and most scientists are also very interested in using data produced by others for their own work. However, studies show that there is still widespread hesitancy to share data, mainly because of fears that the data producer will not receive proper acknowledgement and credit for the original work.

These apprehensions become stronger when discussing more "complex" data containers – how to give "proper" credit if only parts of an aggregated dataset, or a collection of datasets, were actually used in later scientific works? Indeed, many scientists deem it inappropriate or misleading to attribute "collective" credit to everyone who contributed to a collection.

Proposed solutions, now under investigation by various projects focus on two approaches: 1) making the attribution information supplied together with datasets both more detailed and easier to interpret for end users; and 2) providing means for data centres and RIs to extract usage statistics for collection members based on harvested bibliometrics information available for the collections. The first of these could be achieved by e.g., labelling every individual datum with a code indicating the producer, or minting PIDs (DOIs) for the smallest relevant subsets of data, e.g., from a given researcher, group or measurement facility. Based on such information, a data end user can provide detailed provenance about datasets used (at least in article text). The second approach may combine tracing downloads and other access events at the data centre or repository level with bibliometrics, with the aim to produce usage statistics at regular intervals or on demand (from a data producer). However, handling each file's records individually would quickly become cumbersome, so methods of reliably identifying groups of files should be considered.

A particular challenge is to identify and collect references to data objects that are present inside of other data objects, or result from actions of software and workflows. This is particularly true as many of these references will be dynamically created as a range of data is accessed, e.g., as a result of queries like "return the temperature time series for each day over the last decade for station X", taking into account that the actual instrumentation will have changed during the period and/or had periods of non-production. The data extraction may be handled succinctly by a workflow that starts by accessing an explicitly identified catalogue (another dynamic data object) and then traversing the available instruments for each time step before requesting the observational traces (or more likely derivatives of those traces). Such a workflow could be configured to collect, (re-)format and pass on due credit information related to the accessed primary data from stored provenance data, but that may not be accessible, or may be summarised or truncated, because of volume constraints.

- Organisation of (RI-operated) metadata systems that will allow fast and flexible bibliometrics data mining and impact analysis.

Sources: [Socha 2012], ePIC and DataCite PID workshop (Paris, 2015)[4], Make Data Count project, Crossref, OpenAIRE, THOR.

By analysing information about the usage of research data, e.g., through collecting citations and references from a variety of (academic) sources, it is possible to extract interesting knowledge of e.g., what (subsets of) datasets are of interest, who has been accessing the data and how, and in what way they have been used and for what purpose.

Traditionally, this data usage mining is performed based on searching through citation indices or by full-text searches of academic literature (applying the same methods as for articles, e.g., Crossref, Scopus, Web of Science), sometimes also augmented by counting downloads or searches for data at repositories and data portals. However, up till recently, citations of datasets were not routinely indexed by many publishers and indices, and such services are still not comprehensively available across all science fields. At least partly, this is due to limits in the design of citation record databases and the insufficient capacity of lookup services. Here, updated technologies and increased use of, e.g., semantic web-based databases, should bring large improvements.

However, it is important to cover also non-traditional media and content types. Such "altmetrics sources" include Mendeley, CiteULike and ScienceSeeker, as well as Facebook and Twitter. Indeed, while references to research data (rather than research output) in social media may not be very common in Earth Science yet, it may become more prevalent, e.g., where inferences from digital-media activity complement direct observations in poorly instrumented regions. (There are already examples from e.g., astronomy.) Data are in any case already being referred to in many other forms of non-peer-reviewed science-related content, such as Wikipedia articles, Reddit posts, and blogs. Since authors using these "alternative" information outlets are less likely

to use PIDs or other standard citation formats, it is a great challenge to bibliometrics mining systems to identify and properly attribute such references.

- Discovery and sharing, especially of data contained in "complex data objects", may be enhanced by the use of data type registries that facilitate subset identification (and retrieval).

Sources: RDA Data Type Registries Working Group, EUDAT

Data sharing requires that data can be parsed, understood and reused by both people and applications other than those that created the data. Ideally, the metadata will contain exhaustive information about all relevant aspects, e.g., measurement units, geographical reference systems, variable names, etc. However, even if present, such information may not be readily interpretable – it may be expressed in different languages, or contain non-standard terminology. There is a need for a support system that allows for a precise characterisation of the parameter descriptions in a way that can be accessed and understood by both human users and machine-actionable workflows.

Registries containing persistently and uniquely identified Data Type definitions offer one solution that is highly configurable and can be adapted to needs of specific scientific disciplines and research infrastructures. In addition to the basic properties listed above, the type registry entries can also contain relationships with other types (e.g., parent and child, or more complex ones), pointers to services useful for processing or interpretation, or links to data convertors. Data providers can choose to register their own data types (possibly using their own namespace), apply definitions provided by others, or apply a mix of these approaches. The PIDs of the applicable data types are then inserted into the data objects' metadata, and can also be exposed via cataloguing services and search interfaces.

The RDA Data Type Registry working group has designed a prototype registry server, which is currently being tested by a number of RIs and organisations. In a second phase, the RDA group will continue the development of the registry concept by formulating a data model and expression for types, designing a functional specification for type registries, and investigating different options for federating type registries at both technical and organisational levels. The adoption of unambiguous and clear annotation of data, as offered by Data Types, should go a long way towards allaying researchers' concerns that their data will be "misused", either in an erroneous fashion, or for inappropriate purposes.

5.1.3 A longer term horizon

As discussed in a recent report from the RDA Data Fabric Interest Group [Almas 2015], both the increasing amounts of available data and the rapidly evolving ecosystem of computing services, there will have to be an intensifying focus on interconnectedness and interoperability in order to make best use of the funding and resources available to scientists (and society). Tools and technologies including cloud-based processing and storage, and increasing application of machine-actionable workflows including autonomous information searches and data analyses, will all rely on sustainable and reliable systems for identification and citation of data.

Based on this, we have identified a couple of likely trends for the period up to the year 2020:

- A move towards automation of those aspects of the research data lifecycle that will involve basic tasks like assigning identifiers and citing or referring to all kinds of resources – including data and metadata objects, software, workflows, etc.
- Evolution towards more complex "collections" of research resources, like Research Objects, that will necessitate more flexible approaches towards both strategies for identification and detailed, unambiguous citation or referencing parts of such objects.

Much more tightly integrated systems for metadata, provenance, identification and citation will evolve (pushed by data producers, publishers and data centres), offering rapid and trusted feedback on data usage and impact.

5.2 Summary of analysis highlighting implications and issues

Tools and services now under development that will allow seamless linking of data, articles, people, etc. are likely to have a large impact on individual researchers, institutions, publishers and stakeholders by allowing streamlining of the entire data management cycle, virtually instantaneous extraction of usage statistics, and facilitation of data mining and other machine-actionable workflows.

While DOIs for articles, and ORCID identifiers for researchers, are now an accepted part of the scientific information flow, publishing of data may not even consider identifiers for other resources (except for publications, for which DOIs are well established). To speed up the adaptation, both current and future technologies for (data) identification and citation must not only be flexible enough to serve a wide range of existing research environments, but they also have to be shown to provide clear benefits to both producers, curators and end users.

Indeed, while some research communities and infrastructures have fully embraced the consistent use of PIDs for data, metadata and other resources throughout the entire data lifecycle, many others are only beginning to think about using them. Important reasons for this hesitancy or tardiness include a substantial knowledge gap, perceived high investment costs (both for personnel, hardware and software), and a lack of support from the respective scientific communities to change engrained work practices.

ENVRIplus is expected to play an important role in defining best practices for first applying identifiers to data and other research resources – including the researchers themselves – and secondly, how use them for citations and provenance tracking. This will be achieved by 1) designing and building demonstrators and implementations based on concrete needs and requirements of ENVRIplus member RIs; and 2) providing documentation and instructional materials that can be used for training activities.

From a technological point of view, there are two pervasive challenges faced by all those who are engaged in stages of the data lifecycle or are using or producing data in their research or for decision support. One, there are diverse suggestions, but not agreed and widely adopted standards, underpinning the necessary actions, whether those actions are carried out by humans or software. Two, today there aren't good tools and technologies that make it easy for humans or software to perform these tasks efficiently.

There is a great deal of work underway, and we can be optimistic about viable deployable support for data identification and citation becoming available within the next few years. This poses another two challenges: 1) how to identify and align with the software and methods that will be most widely supported and adopted; and 2) how best to use the emerging software, metadata standards and proposed methods in the ENVRIplus context. That requires developing standard practices, metadata and protocols that allow interworking within and between the RIs and other organisations. Indeed, cataloguing, curation and provenance all need to make effective use of the functionality and facilities data identification and citation will provide. Conversely, the work on catalogues may provide facilities for PID registries with associated metadata. The basic consistency for data identification and citation should be achievable within the ENVRIplus project's lifetime. But finding ways of succinctly, efficiently and precisely identifying the growing volumes and many subtleties of the data used by and produced by future data-driven research will always be a challenge.

There are further considerations that may be addressed in the future. These are enumerated here in no particular order:

1. Identification and cataloguing: The relationship between data identification and cataloguing is very close.
2. Roles for data identifiers: It is obvious that cataloguing and provenance will need reliable data identifiers to refer to data from their records.
3. Raising the level of discourse: Developing precise, abstract models, then clarifying them through discussion and revision is of greater and long-term benefit. The ENVRIplus Reference Model provides a vocabulary and context where such discussion takes place.
4. Accommodating diversity: Although the campaign for harmonisation [e.g. of cataloguing standards and data models] is vital, it will never completely achieve conformity. [To alleviate resulting problems, one can offer] functions that map, sometimes with loss of precision, to a common interchange form, [as well as introduce] fields in the data-identity registry capable of holding any user defined material.
5. Registry platforms: It is necessary to build registries and other catalogues on top of high-quality database systems.
6. Temporal patterns: Registries as well as catalogues can be built continuously or incrementally, and the effects of these strategies need to be considered carefully.
7. Distribution patterns: In order to handle users' transient and initial interactions between e.g. a catalogue, a PID resolver and a data store, while still promoting persistent complete records to authority sites, replication and/or federation between registration and resolution servers may be needed.

6 MAPPING THE PID PROVIDER, PUBLISHER & INDEXER LANDSCAPE

For RIs and researchers who are getting ready to design and implement strategies and practices for making persistent identifiers for data (and other resources) an integral part of their research data life cycles, the diversity of PID service providers, publisher requirements and data citation indexing possibilities can be quite complex.

In this Chapter, we have attempted to map out the landscape of PID service providers, publishers and indexing organisations that have the highest relevance and importance to environmental research infrastructures and their user communities. We have also looked at some global research data-oriented initiatives that are of importance to ENVRIplus members.

6.1 PID providers

6.1.1 ePIC — the European Persistent Identifier Consortium

The European Persistent Identifier Consortium (ePIC; see http://www.pidconsortium.eu) provides PID services to European centres that store research data. Consortium members are: IT Centre for Science (CSC) in Finland, DKRZ in Germany (see **Chapter 7.1.2** for a description), Greek Research and Technology Network (GRNet) in Greece, Gesellschaft für wissenschaftliche Datenverarbeitung Göttingen (GWDG) in Germany, Centre for High-Performance Computing at the KTH (PDC KTH) in Sweden, and SURFsara in the Netherlands.

The ePIC PIDs are based on handles and can be used for data objects and data collections as the users of the services see fit. The consortium members share services and APIs, and promises that if one centre is out of order, the other centres can still resolve PIDs [ePIC 2016a]. Services offered for ePIC PIDs are generation, resolution and replication of PIDs, and a mirror of the Global Handle Server in Europe, which ensures that PIDs can be resolved even when parts of the global network are down [ePIC 2016b].

6.1.2 DataCite

DataCite (see http://datacite.org) is a non-profit, community-driven organisation that provides DOI services for research data, including generation and allocation of DOIs and related metadata,

and discovery services for research data [DataCite 2016a]. The services are offered in collaboration with its community of members, which includes data centres, libraries, universities and other organisations [DataCite 2016b]. DataCite also works on advocacy for data citation through events and activities in research communities [DataCite 2016a].

In collaboration with ORCID, DataCite offers the service DataCite Profiles which connects researchers' ORCID records and can update researchers' ORCID records for works that have DOIs [DataCite Profiles 2016]. The Citation formatter is another collaboratively developed service (with Crossref (see **Section 6.3.1** for a description), mEDRA https://www.medra.org/ and ISTIC http://www.doi.org.cn/portal/index.htm) which allows programs to create citations by extracting metadata from DOI records. Currently, over 5000 citation styles in 45 languages are supported [DataCite 2016c]. Connected to metadata, the service OAI-PMH Data Provider facilitates harvesting of metadata from resources with a DataCite DOI [DataCite 2016d]

A new collaboration between DataCite, ORCID and Crossref on organisational PIDs, the Organization Identifier Project, was presented at the PIDapalooza conference on 10 November 2016. For information about the project described from the member organisations' different contexts, and for links to the documents, see DataCite: [Fenner 2016b], ORCID: [Haak 2016], and Crossref: [Pentz 2016].

6.1.3 ORCID — Open Researcher and Contributor ID

Open Researcher and Contributor ID (ORCID; see https://orcid.org/) is a non-profit organisation that provides PIDs to uniquely identify researchers and to connect researchers to their works and research outputs [ORCID 2016a]. ORCID's mantra is "enter once, re-use often" [ORCID 2016b], meaning that data in a researcher's ORCID record can be imported into services provided by other actors, e.g. funders, publishers and libraries. The ORCID tools, and other work done by the organisation, are open and transparent, and are developed in close collaboration with the research community [ORCID 2016a]. ORCID IDs are used in the research workflow (e.g. submission of grants and manuscripts) by its member organisations as well as non-members. Both types of users include actors such as funders, research institutions, libraries and repositories [ORCID 2016c]. In November 2016 there were 2,763,353 ORCID IDs [ORCID 2016d].

The ORCID ID distinguishes researchers. It is a unique and persistent identifier, governed by ORCID. ORCID IDs are actionable, meaning that they can be resolved to obtain information about the person identified by the ID. In contrast to a person's name, the ORCID ID is unambiguous. In contrast to a researcher's email, the ID is persistent. In addition to providing digital identifiers for researchers (and contributors more generally) ORCID also facilitates the integration of IDs in research workflows as well as linking of researchers with research artefacts, such as datasets.

ORCID collaborates with other PID organisations, such as DataCite and Crossref, to create tools for linking research objects and people. For example, the Auto-Update service makes it possible to automatically update a researcher's ORCID record when new publications are published (if the researcher gives permission to link these services) [ORCID 2016e]. This simplifies the task of individual researchers keeping their publication lists up to date.

6.1.4 ISNI — International Standard Name Identifier

International Standard Name Identifier (ISNI; see http://www.isni.org/) is an ISO certified standard for identification of creators of works, including researchers, publishers, producers, artists, and inventors to enable unambiguous attribution to a work's creator. ISNI International Authority (IA) is the organisation behind the ISNI PIDs, and is a non-profit organisation consisting of members of rights management organisations and libraries [ISNI 2016b]. ISNI is a so called "bridge-identifier" linking domains together, and is used for linked data applications. In November 2016 there were 8.55 million names in the ISNI database, of which 2.58 million are researchers [ISNI 2016a]. ISNI is part of a set of international standard identifiers, including the DOI standard, ISBN (International Standard Book Number) for books, and ISSN (International Standard Serial Number) used for identifying e.g. journals, collections, databases [ISNI 2016b].

The differences between ISNI and ORCID are related to their different foci and user communities: whereas ISNI can identify anyone responsible for a work, and have many users in creative arts, ORCID focuses specifically on researchers producing scholarly output, and the stakeholders involved in research. The organisations coordinate their work to ensure there is no overlap in assignment of PIDs, and ORCID follows the ISNI ISO standard [ORCID 2016f]. The management of historical and deceased persons differ between ISNI and ORCID. ISNI allows for creation of PIDs for persons who cannot themselves create the PID, whereas ORCID requires each person to create their own ORCID ID [Jessop 2016].

6.1.5 Others

A growing number of general purpose data repositories issue DOIs for published datasets.

Dataverse (http://dataverse.org) is an open source web application and data repository which allows researchers to "share, preserve, cite, explore, and analyse research data" and is a collaboration between the Institute for Quantitative Social Science (IQSS) Harvard University Library, and Harvard University Information Technology [Dataverse 2016a]. It is open to researchers in all disciplines throughout the world [Dataverse 2016b]. Both DOIs and Handle system PIDs are supported [Dataverse 2016c], and can be provided by either DataCite or EZID (http://ezid.cdlib.org/ — a service for minting DOIs and ARKs offered by the California Digital Library) [Dataverse 2016d]. To encourage correct data citation, Dataverse generates bibliographic references for datasets, based on the Joint Declaration of Data Citation Principles [FORCE11 2014a, Dataverse 2016e] (see **Section 2.4.1**).

figshare (https://figshare.com/) is a repository for sharing a wide range of research output types, including datasets, software code, and posters [figshare 2017a]. The business model of figshare "supports sustainability" with a guarantee of 10 years of persistent availability for uploaded resources [figshare 2017b] by using the CLOCKSS Archive's network of scholarly publishers and research libraries to archive scholarly resources published on the web [Hahnel 2012]. figshare uses DataCite DOIs via the California Digital Library for uploaded resources. For institutional clients, figshare has an agreement with DataCite to connect clients to DataCite nodes in order to set up minting of DOIs by the institutions, but it is also possible to request that figshare mint DOIs if an institution for some reason does not wish to mint their own DOIs [Hahnel 2015]. As a complement to minting DOIs to uploaded resources, figshare allows users to reserve DOIs for resources that are not ready to be published and publicly available yet [figshare 2017c]. In the spirit of linked open data, GitHub and figshare accounts can be connected in order to facilitate finding data and the software that has been used to generate process, or analyse the data, and might be needed to understand and use the data. The linking of the two repositories was an early initiative to link resources stored in different repositories, and was a collaborative project between GitHub, Mozilla Science Lab and figshare [Hahnel 2014].

ResearchGate (http://researchgate.net), the social networking site for researchers, offers members to generate ResearchGate DOIs for unpublished work [ResearchGate 2016a], for example raw data and research proposals [ResearchGate 2016b], but not for published articles and books as these in most cases already have a DOI. ResearchGate DOIs are issued by DataCite [Koshoffer 2014]. As Koshoffer [2014] notes, ResearchGate allows for content with ResearchGate DOIs to be deleted, which contradicts the notion of persistence for resources that have been issued a PID.

Zenodo (https://zenodo.org/) is an open science platform and data repository developed within the OpenAIREplus project, with CERN as the major developer. Publishing data in Zenodo is free for the long tail of science [Zenodo 2016a], allowing for up to 50 GB per dataset [Zenodo 2016b]. Zenodo issues DataCite DOIs and supports harvesting of content, including metadata of datasets with the OAI-PMH protocol [Zenodo 2016c]. Similarly to figshare, Zenodo offers a link between users' Zenodo accounts and GitHub accounts, to facilitate finding the data and code that have been, or should be, used together.

In addition, there are various services to assign PIDs (handles) to web links — these can be used to identify e.g. documents not yet in a repository, images and visualizations (especially those created by passing parameters embedded into the URL of a web-based tool). One example is ShortRef (http://www.shortref.org/).

6.2 Associations and societies for scholarly publishers

As part of the preparations for the negotiation rounds described below in **Chapter 9,** it is important to first identify the associations and societies for scholarly publishers that are the most important ones for publishing research based on data provided by ENVRIplus members. The associations and societies for scholarly publishers often have a mandate from their members to drive the technology development for scholarly publishing. Here we list some of the relevant ones. Individual publishers, or even journals, may be relevant to invite to a negotiation session. This will be followed up on throughout the WP6 project.

6.2.1 Associations and societies for scholarly publishers

The prevalent associations and societies for scholarly and professional publishers are described below.

Association of Learned & Professional Society Publishers, ALPSP (http://www.alpsp.org/) is a trade organisation for non-profit scholarly publishers and associated organisations that support publishers. Out of ALPSP's 332 members, 208 are non-profit publishers, 50 are organisations providing services for publishers, and 36 are commercial publishers [ALPSP 017a].

International Association of Scientific, Technical and Medical (STM) Publishers (http://www.stm-assoc.org/) is a trade association which includes over 120 academic and professional publishers. The members of the association publish approximately 66% of all academic journal articles published yearly [STM Publishers 2017a]. Members include the biggest scholarly publishers Elsevier, Sage, Springer, Taylor & Francis and Wiley, as well as smaller publishers. The association works actively with standards and technology development, and has been involved in the development of Scholix, Scholarly Link Exchange (http://www.scholix.org/home), see Chapter 7.2.2 below.

Open Access Scholarly Publishers Association, OASPA (http://oaspa.org/) represents the interests of about 90 Open Access journal and book publishers across all disciplines [OASPA 2017a]. OASPA is active in promoting standards for of Open Access scholarly communication [OASPA 2017b] and collaborates with Crossref on the DOI Event Tracker project, at the same time encouraging publishers of all types to make reference lists freely available through Crossref's services [Redhead 2015]. Members include Copernicus Publications that publishes over 20 Open Access journals in the earth sciences [Copernicus 2017], Public Library of Science, PLoS (https://www.plos.org/), Cambridge University Press, Hindawi Publishing Corporation, Oxford University Press, SAGE Publishing, Springer Nature, and Taylor & Francis [OASPA 2017c]. Members of OASPA are required to use liberal licenses for publications, with CC-BY being highly recommended [OASPA 2017d]

6.3 Indexing agencies and services

Here we list some of the indexing agencies and services that are of relevance to producers and publishers of environmental data.

6.3.1 Crossref

Crossref (http://crossref.org) is a DOI Registration Agency under the International DOI Foundation, and is run as a non-profit organisation through the Publishers International Linking Association (PILA) [Crossref 2015a]. In October 2015, Crossref had 5322 members of the type publishers and societies [Crossref 2015b], and 1976 libraries [Crossref 2015c]. The core services include providing DOIs and metadata for digital academic works and linking between the participating publishers' digital publications. (This covers both inbound links, i.e. to a journal

article, and outbound links, i.e. from a journal article's reference list to each cited publication.) The DOI resolution system that enables look-up of DOIs is the Handle.Net Registry (http://www.handle.net/), managed by the DONA Foundation (https://www.dona.net/). Publishers who wish to use the Crossref services agree to submit a basic set of metadata about their articles according to the Crossref scheme: journal title, ISSN, first author, year, volume and issue, page numbers, DOI and URL, with the possibility of adding more metadata as they see fit [Crossref 2013]. As mentioned, Crossref is involved in several collaborations with other PID providers, such as DataCite and ORCID, to enable linking between different types of research objects, including datasets and people (see **Chapter 6.1.2** for DataCite and **Chapter 6.1.3** for ORCID).

6.3.2 Web of Science and Data Citation Index

Web of Science (http://ipscience.thomsonreuters.com/product/web-of-science/ previously known as (ISI) Web of Knowledge is an online subscription-based scientific citation indexing service. It was previously a product developed and maintained by Thomson Reuters Intellectual Property & Science, but is, since autumn 2016, managed by Clarivate Analytics, a company owned by Onex Corporation and Baring Private Equity Asia [PR Newswire 2016]. Web of Science provides a comprehensive citation index for scientific publications in a broad range of disciplines from the sciences, arts and humanities, and the social sciences, and is a widely used tool for bibliometrics analysis. Publication types indexed by Web of Science include journals, books, and conference proceedings. A related product, Data Citation Index, indexes data from a wide range of disciplines published in data repositories all over the world. The metadata of the datasets are linked to the publications indexed in the Web of Science which enables discovery of results and data in the same search environment [Web of Science 2016a]. The Data Citation Index uses the DataCite metadata standard for data citation is used for Data Citation Index because of its general acceptance and the possibilities of describing many types of data from many disciplines [Web of Science 2016b].

The Data Citation Index manages three types of resources: data repository, data study, and dataset. A data repository is defined as "a database or collection comprising data studies, and datasets which stores and provides access to the raw data" [Web of Science 2016b]. These can consist of a data study, which is a "description of studies or experiments held in repositories with the associated data which have been used in the data study" [Web of Science 2016b]. A dataset, finally, is defined as "a single or coherent set of data or a data file provided by the repository, as part of a collection, data study or experiment" [Web of Science 2016b].

6.3.3 Mendeley

Mendeley (http://www.mendeley.com) is a reference management service and social network for the scholarly community, and includes a citation plugin that facilitates citation when using word processors [Mendeley 2017a], similarly to EndNote. Mendeley also offers sharing datasets, and discovery of published datasets, in a cloud-based repository feature which includes the possibility to assign DOIs to datasets, versioning support, and linking articles and datasets [Mendeley 2017b]. Mendeley was developed as open source software and was acquired by one of the major academic publishing companies, namely Elsevier, in 2013, which has been severely criticised in the open access and open data communities due to Elsevier's profit margins and paywalls [Amirtha 2015].

6.3.4 Altmetrics

Altmetrics are alternative metrics of scholarly impact that move beyond the traditional measures based on citation counts, such as journal impact factor and most-cited papers. Examples of alternative metrics can be categorised by views, discussions (including tweets and shares on Facebook, comments on journal web sites and Wikipedia mentions), citations, recommendations and saves. Altmetrics also facilitates measuring impact for a wider range of research objects and stakeholders, such as datasets, presentations, subsets of publications, funders, and researchers

[Lin 2013]. There is a plethora of actors providing altmetrics, including altmetric.com, Plum Analytics, ImpactStory, and PLoS. The usefulness of altmetrics is debated. An extensive study of altmetric.com indicators has confirmed previous studies showing that there is a positive correlation between altmetrics and citations, be it a weak one [Costas 2015]. The study also showed that the use of altmetric.com indicators (which is the most widely used set of metrics) were only about 15-24% of published papers that had an altmetrics activity (score) related to them, mostly within the humanities, social sciences, life sciences and medicine [Costas 2015].

6.3.5 Making Data Count

The Making Data Count project (http://mdc.lagotto.io/), active between 2014 and 2015, had the aim to develop a reference model for data-level metrics and was a collaboration between the California Digital Library UC3, PLoS and DataONE [MakingDataCount 2015a].

The survey conducted within the project showed that researchers found the number of citations to be the most important metric for data, followed by number of downloads and views, the lower ranked metrics being the more easily implemented, and keeping track of citations being the most complex for data repositories to handle [Kratz 2015].

The open source software Lagotto (http://lagotto.io/) was another outcome of the project, facilitating tracking of events (such as views, downloads and mentions) related to research objects, including scholarly publications and data [Lagotto 2017]. The DataCite Event Data service includes an API to import data into Lagotto [DataCite 2017].

6.3.6 Bibliographic databases

Bibliographic databases index scholarly journal publications including journal articles, books, and conference proceedings and provide rich bibliographic information about the indexed works, including abstracts, keywords and provenance information. The largest bibliographic databases are Scopus (https://www.elsevier.com/solutions/scopus) which indexes 21,500 scholarly journals, Web of Science (http://ipscience.thomsonreuters.com/product/web-of-science/), and Google Scholar estimated to contain about 160 million documents in May 2014 [Wikipedia 2017b].

6.3.7 Discovery services

Discovery services for libraries gives access to different types of content in one service, for example content from bibliographic databases, full-text publications that the library purchases access to, and the library's local catalogues and collections. This provides a seamless search and discovery interface for users. Content is linked with a link resolver as a hub, often based on the OpenURL format (also an ANSI/NISO standard: Z39.88-2004), which enables linking of bibliographic citations to the full-text publication [Wikipedia 2017c]. The SFX software is the most commonly used OpenURL resolver by libraries and scholarly publishers [Wikipedia 2017d]. Discover system solutions used by libraries include WorldCat Discovery delivered by OCLC, Primo delivered by Ex Libris, 360 Link delivered by ProQuest, Full Text Finder delivered by Ebsco.

6.4 Global research data-oriented initiatives

There exist a number of global organisations that aim to bring together individuals and communities interested in developing and promoting new services and practices for efficient research data management. Here we describe two of them, the Research Data Alliance and CODATA.

6.4.1 Research Data Alliance (RDA)

Research Data Alliance (RDA) is a community-driven organisation with the aim to facilitate data sharing and development of the necessary social and technical infrastructure to support sharing of data. It was created in 2013 by the European Commission, the United States Government's National Science Foundation and National Institute of Standards and Technology, and the

Australian Government's Department of Innovation. In November 2016, RDA had 4500+ members from 115 countries [RDA 2016a], representing a wide variety of professions and expertise, including scientists, technologists, and librarians. RDA Europe is a project that focuses on the European context by organising workshops, training events and supports travel to RDA activities [RDA Europe 2016].

RDA is built up from a large number of groups focussing on different aspects of data and infrastructures. The work in these groups is driven entirely by the individual members, who are all volunteering their time and expertise. There are two types of groupings, with different scales in scope and time: Open-ended Interest groups (IGs), and time-limited (typically 18 months) Working Groups (WGs). The IGs focus on finding solutions to more over-arching data sharing problems and topics, including metadata, provenance and data publishing. In contrast, the working groups (WGs) are set up with the aim to investigate more well-defined topics, for example developing an API for data collections or finding recommendations for dynamic data citation. At the end of their allocated time, WGs are expected to present their results as recommendations or demonstrators. In January 2017 there were 51 IGs and 35 WGs listed on the RDA website [RDA 2016b, RDA 2016c].

The following RDA groups are of direct interest to ENVRIplus Work Package 6 activities. Currently, all these groups have at least one member representing an ENVRIplus partner.

Interest groups:

- Data Fabric IG (https://www.rd-alliance.org/group/data-fabric-ig.html)
- Data Foundations and Terminology IG https://www.rd-alliance.org/groups/data-foundations-and-terminology-ig.html
- Data Versioning IG (starting) https://www.rd-alliance.org/groups/data-versioning-ig
- Marine data Harmonization IG https://www.rd-alliance.org/groups/marine-data-harmonization-ig.html
- Metadata IG https://www.rd-alliance.org/groups/metadata-ig.html
- PID IG https://www.rd-alliance.org/groups/pid-interest-group.html
- RDA/WDS Data Publishing IG https://www.rd-alliance.org/groups/rdawds-publishing-data-ig.html

Working groups:

- Data Citation WG https://www.rd-alliance.org/groups/data-citation-wg.html
- Data Description Registry Interoperability (DDRI) WG https://www.rd-alliance.org/groups/data-description-registry-interoperability.html
- Data Type Registries WG (1 and 2) https://www.rd-alliance.org/groups/data-type-registries-wg.html
- PID Information Types WG (finished) https://www.rd-alliance.org/groups/pid-information-types-wg.html
- PID Kernel Information WG (starting) https://www.rd-alliance.org/groups/pid-kernel-information-wg
- RDA/WDS Publishing Data Bibliometrics WG https://www.rd-alliance.org/groups/rdawds-publishing-data-bibliometrics-wg.html
- Research Data Collections WG https://www.rd-alliance.org/groups/pid-collections-wg.html

6.4.2 CODATA

The Committee on Data for Science and Technology (CODATA) was formed in 1966 by the Scientific Committee of the International Council for Science (ICSU) to "promote and encourage the compilation, evaluation and dissemination of reliable numerical data of importance to science and technology" [CODATA 2016a]. In promoting accessibility and high quality data, CODATA arranges conferences and workshops on topics related to research data management, focussing in particular on promoting cross-disciplinary use of data and on issues that many

disciplines have in common. CODATA has a global perspective, but there are also national committees that arrange national activities. A number of task groups and working groups focus on specific topics, including coordination of multinational data projects, Supplying information on sources of reliable data, and Establishment of format standards to promote data exchange, sharing and compatibility.

The task group CODATA-ICSTI Data Citation Standards and Practices has published two reports: "For Attribution: Developing Data Attribution and Citation Practices and Standards" [Uhlir 2012] published in 2012 and "Out of Cite Out Of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data" [Socha 2013] published in 2013. In 2014, CODATA gained membership of the Group on Earth Observations' (GEO) Data Sharing Working Group [CODATA 2016b].

7 IDENTIFICATION & CITATION FOR RESEARCH DATA: WORK IN PROGRESS

This Chapter presents examples of on-going work related to Identification & Citation that is currently being undertaken in parallel to Work Package 6, both by ENVRIplus partners and various other parties and organisations.

7.1 Parallel activities by ENVRIplus partners

Apart from the activities defined in WP6 description of work, several of the ENVRIplus partners are involved in parallel activities that relate to developing or operating services that involve data identification and citation.

7.1.1 University of Bremen (PANGAEA)

PANGAEA — Data Publisher for Earth & Environmental Science (http://www.pangaea.de/) — is a digital data library and a data publisher for Earth system science. PANGAEA is hosted by the Alfred Wegener Institute for Polar and Marine Research (AWI), Bremerhaven and the Centre for Marine Environmental Sciences (MARUM), Bremen in Germany. It is used as a data repository by various publicly funded international research projects, and by the World Data Centre for Marine Environmental Sciences (WDC-MARE) as a long-term archive. PANGAEA was initially developed in 1987 and has been operational on the Internet since 1995 [Wikipedia 2017e].

Scientific data are archived with related meta-information in a relational database (Sybase) through an editorial system. Open Access data are distributed in standard formats through web services on the Internet, via search engines and portals. Dataset descriptions (metadata) are conforming to the ISO 19115 standard but can be served in various formats (e.g. Directory Interchange Format, Dublin Core) via the OAI-MPH and Web Catalogue Service protocols. They include a bibliographic citation and are persistently identified using Digital Object Identifiers (DOI). Identifier provision and long-term availability of datasets via library catalogues is ensured through cooperation with the German National Library of Science and Technology (TIB). Retrieval of datasets is provided through a full text search engine (based on Apache Lucene/panFMP) [Schindler 2008]. In addition, a data warehouse is operated to provide efficient data compilations [Wikipedia 2017e].

In parallel to contributing to activities related to data identification, which is the main focus of this WP, University of Bremen (PANGAEA) has proposed to discuss and coordinate the identification of researchers and contributors to datasets curated by ENVRIplus RIs, specifically using ORCID.

Following the recent integration of ORCID in PANGAEA, it is now proposed to integrate ORCID in the data management of environmental and Earth science RIs. Research infrastructures involve both technical and social entities. The integration of ORCID in RIs will enable the identification of social entities (scientists, technicians, data managers, etc.) and the linking of social with technical

entities (datasets, software, articles, devices, stations). Such linking will enable unambiguous attribution of the contributions by social entities to the development, implementation, and maintenance of technical entities. See **Chapter 7.2.1** for more details.

As a related activity, PANGAEA has discussed the persistent identification of instruments, as well as platforms and deployments, at PIDapalooza 2016, Reykjavik, Iceland, November 9-10 [Stocker 2016]. The session motivated why it could be important to persistently identify these entities and preserve metadata about them. PANGAEA presented two existing infrastructures that utilise DOI to identify platforms (seismic station networks) and deployments (cruises). The persistent identification of these entities extends the application of PID beyond articles, datasets, and people.

7.1.2 DKRZ (CMIP6 project)

Deutsches Klimarechenzentrum (DKRZ), English name German Climate Computing Centre, is a computing centre that offers services to climate science. Services include high performance computing (HPC) resources, tools for data management and visualization, and consultancy [DKRZ 2016].

Introduction

Presently, climate modelling projects such as CMIP5 or CMIP6 incur millions of files aggregated at various levels (time-series, experiments…). To identify and track them, registered ePIC PIDs are an appropriate means but complex technological and organizational efforts are required to be supported in international data federations like ENES (European Network for Earth System Modelling, see https://verc.enes.org/community/about-enes) and ESGF (Earth System Grid Federation; see http://esgf.llnl.gov/). They are, e.g., included in the CMIP6 file headers and registered as part of the file publication in the data federation. The organizational aspects are coordinated in the World Climate Research Program (WCRP) infrastructure panel (WIP). The technological aspects are implemented in the context of the Earth System Grid Federation (ESGF) and include a highly scalable PID registration infrastructure based on message queues. For citation of these data, aggregated groups of these files (sets of experiments) are equipped with DOIs registered by DataCite.org.

ePIC PIDs: As a member of the PID Consortium ePIC ("Persistent Identifiers for eResearch") DKRZ is involved in development and maintenance of ePIC's PID infrastructure to support the use of long term identifiers for scientific data objects.

DataCite DOIs: A service for the integration of research data into scientific publications has been developed at DKRZ together with the Technical Information Library (TIB, Hannover) and is offered to the scientific Earth System community. DOIs (Digital Object Identifier) and citation codes are provided and registered with DataCite as DOI registration agency. The editorial process includes checks of metadata and research data quality in collaboration with the corresponding data producers.

Persistent Identifiers in the workflow

Table 1 below shows a comparison between the characteristics and suitability of ePIC PIDs and DataCite DOIs for data objects.

ePIC PIDs: For the big model Intercomparison project CMIP6 (https://www.wcrp-climate.org/wgcm-cmip/wgcm-cmip6) it has been decided to integrate handle system based PIDs into the header of every file and integrate the PID registration process into the ESGF file publication process. To tackle the scalability problems of the Handle system with respect to PID registration a distributed message queuing based solution is being deployed and integrated with the ESGF infrastructure. Further use of PIDs in the workflow is under development, e.g. in the context of managing data errata information.

TABLE 1. COMPARISON BETWEEN USE OF EPIC PID AND DATACITE DOI AT DKRZ.

	ePIC PID	DataCite DOI
Data Granularity	Lowest granularity, file level; higher granularities possible by aggregation to PID Collections	Citation granularity
Main (planned) use	Data handling (internal) and data identification (internal & external) as well as data annotation (e.g. errata information etc.)	Data citation followed by data access/identification
Data Quality	CMIP project compliance checks (file metadata, controlled vocabularies), data objects "as is"	Multiple documented checks
Data Quantity	high numbers of single files, high degree of automation	moderate number of DOI, low degree of automation, feedback to data producer possible
Operational capability	Pilot applications running, integration into international ESGF data federation, scalability solution for PID registration. Operational deployment in 2017	Operational
Main affected work fields	Projects CMIP, EUDAT	CMIP, CORDEX, single customers & partners

DataCite DOIs: For identification and citation, most long term archived (LTA) data are assigned a DOI and registered at DataCite. Here two main challenges had to be solved. First, unlike a printed paper, stored data often undergo subsequent improvements or corrections for which the identifier system has to account. Here the data dissemination system has to account for versioning: from the data upload to the identification and to the Graphical User Interface (GUI).

Second, as the upload of the metadata to DataCite involves the completion of final quality checks, the need for data citation in an earlier stage of the workflow arose in some projects (see below Early Citation).

For details of both challenges, see [Stockhause 2015].

Between first in-project publication and minting of the DOI:

Early Citation in the project CMIP6: For data that have a DOI, a high degree of stability is expected which is not the case after early publication in the project – corrections might follow. However, already in this early phase the need for an Early Citation became evident, too.

The initial (default) citation information is collected from the data producer via the scientific contact person given in the file header. It is checked and entered into the meta database (MDB) by the data repository. The information is accessible to GUIs and other repositories by an API. In case citation information changes, the data producers have to update the MDB via a dedicated GUI. A check of the new information follows.

7.1.3 Marine RIs

The ENVRIplus partner Ifremer has recently published a set of guidelines and recommendations for how to implement persistent identifiers for ocean data [Merceur 2016]. The guidelines focus on two important cases: PIDs for dynamic datasets that change over time as new (observational) data become available, and PIDs for sensor platforms, research vessels and measurement campaigns.

Dynamic dataset DOIs

Central to the described approach is the maintenance of an extensive cataloguing system, that can keep track of all datasets and the dates at which data are recorded and made available. The catalogue also contains a wealth of metadata about the data objects, and this allows the creation of information-rich landing pages for data objects. As an example, the global Argo landing pages include the following metadata fields:

- General data (title, author, publication date, description, data license …)
- Link to the Users' Manual
- List of available snapshots (with dates)
- Link to the GDAC FTP site, for downloading the dataset
- Suggested citation
- Geographical area covered
- DOI metadata export link
- List of citing publications (created for Argo by the University of California San Diego)
- List of associated datasets
- Links to social networks

To cope with the frequent updates of datasets, the recommended approach for the involved data centres that are operating marine data repositories is to create and store separately snapshots of the database. These snapshots are however not given individual persistent identifiers. Instead, the datasets are registered (with DataCite) at their creation, and assigned a DOI that they will always keep. The snapshots are instead distinguished by appending a snapshot-specific additional "appendage" that is appended to the end of the dataset DOI, preceded by a hashtag.

The resolving agent at DOI.org will, when it is confronted by an identifier such as http://dx.doi.org/10.17882/42182#46618, redirect to the base landing page of the dataset itself, in this case http://www.seanoe.org/data/00311/42182/. The server that is responsible for the landing pages then recognises that it is also being passed a snapshot identifier, i.e. 46618, which a lookup in the underlying cataloguing database identifies as the Argo GDAC dataset snapshot of November 8, 2016. If no snapshot identifier is appended, the landing page server will display information about the latest snapshot available.

Cruise-specific DOIs

On behalf of TGIR, the French scientific infrastructure for research vessels, a DOI is assigned to each French oceanographic cruise by the Sismer organisation. The DOI should be cited by any author using data from the cruise, which makes it easy to extract information on how often a cruise is cited -- data that the funding agencies are very interested in. In addition, it is possible to look for linkages between cited datasets and cited cruises. Cross references of DOIs between data centres also improves the access to cruise research data. Finally, the cruise DOI also helps to identify the cruise as a scientific activity, and also improves the visibility of the Fleet TGIR as an infrastructure.

As with the Argo datasets, the landing pages of the cruise DOIs also provide rich metadata to the users:

- General data (title, objective, authors with possible links via their ORCID records, …)
- Navigation map
- Citation text suggestion
- Links to external documents and sites
- Videos of underwater vehicles (if applicable and available)
- Published data
- Data managed by Sismer
- Sampling operations

- Diving operations (if applicable)
- Moorings (if deployed)
- Bibliography

7.1.4 ENVRIplus implementation cases

IC-01 — (Dynamic) Data identification and citation (A. Vermeulen *et al.*)

Identification of data (and associated metadata) throughout all stages of processing is really central in any RI. This can be ensured by allocating unique and persistent digital identifiers (PIDs) to data objects throughout the data processing life cycle. The PIDs allow unambiguous references be made to data during curation, cataloguing and support provenance tracking. They are also a necessary requirements for correct citation (and hence attribution) of the data by end users, as this is only precise and easily used when persistent identifiers exist and are applied in the attribution.

At the same time, research data may be changing over time as new records are added, errors are corrected and obsolete records are deleted from a dataset. Researchers rarely use an entire dataset or stream data as it is, but rather create specific subsets tailored to their experiments. In order to keep such experiments reproducible and to share and cite the particular data used in a study, researchers need means for identifying the exact version of a subset as it was used during a specific execution of a workflow, even if the data source is continuously evolving. (Many automated workflows already provide this automatically. However, people rarely use such information at present, and much research also involves human decisions and judgements that are harder to capture and reuse.)

In this implementation case we evaluate the requirements from the RI's gathered on the topics of Identification and Citation, and define the best candidates for technologies that will allow implementation of data citation for dynamic datasets and collections of datasets.

The RDA working group on data citation (https://www.rd-alliance.org/group/data-citation-wg.html) has laid out a solution direction that allows accessing individual subsets of data in a dynamic context, supporting the identification of fine granular subsets of evolving data. This approach centres on assigning PIDs to the actual queries made by users to extract data, rather than to the data objects containing the extracted data. The process is very lightweight and scales with increasing amounts of data. It preserves the subset creation process and thus contributes to the reproducibility of an experiment also on the intellectual level, providing provenance details and metadata.

The RDA recommendation on data citation requires that all metadata -- and possibly also the data -- is stored in the form of a versionable database. While this can be implemented relatively straightforwardly from scratch, a major reconstruction effort is required for existing metadata databases and/or flat file-based data storage approaches. Another major challenge is the requirement to guarantee that the database will be "future proof" and will also work 20 years from now supporting the same queries. Proper attribution also requires that citation services like DataCite support the harvesting of the contributor metadata in their citation indices. Other challenges are that this requires a mechanism to identify the uniqueness of queries and that all data are stored with stable sorting.

IC-09: Quantitative Accounting of Open Data Use (M. Fiebig *et al.*)

The ENVRIplus survey responses showed that a majority of RIs find that it is absolutely necessary to ensure that credit for producing and managing of scientific datasets is "properly assigned", down to the level of individual principal investigators (PIs) in charge of measurement and observation stations. This result is in line with many earlier studies which have shown that the perceived lack of proper attribution of data is a major reason for the hesitancy felt by many researchers to share their data openly.

While there is reasonable confidence that identification and subsequent citation practices will result in adequate possibilities to trace and account for usage of individual datasets, for example by assigning DOIs to data objects, many RIs are apprehensive and concerned about how usage statistics for data collections can be fairly and correctly translated into "usage credit" for data items that are members of such collections — i.e., they fear it will be difficult or even impossible to trace back the provenance of actually used collection items to their individual provider through the currently used data citation practices and bibliometrics tools.

One of the main causes for this concern is the increasing pressure from policy makers and funding agencies towards research groups and organizations to show that their data are not only being released under Open Access policies, but that they are also re-used, thus maximizing the benefit of public investments. Indeed, funding agencies are pushing towards increasingly more open data policies, including re-distribution and commercial use of data. At the same time, the same usage statistics (mainly in the form of "citation metrics"), remains the basis of documenting of scientific merit that is paramount for scientists' employment and stations' funding.

Within ENVRIplus, a decision has been made to set up a task force to investigate different methods for how data usage metrics and statistics can be improved, not only for individual datasets but especially for collections of datasets.

The strategy currently discussed defines two uses of DOIs to achieve the above mentioned objectives:

- **Item DOIs:** used to identify every single dataset contained in the primary archive used by an RI. The granularity needs to be fine enough to identify data down to the contributing scientist in a quantifiable way, but coarse enough to be practical for the data repository considering its data structure. A typical example for data from surface station networks would be one year's worth of data for one instrument to receive one DOI.
- **Collection DOIs**: used to identify a user defined collection of data that may reside in several distributed primary archives in a way convenient for the user. In order to facilitate quantifiable credit to the contributing scientist, the collection DOI needs to refer back to all primary DOIs fully or partially contained in the collection.

This approach allows indexing agencies to resolve a data quotation event down to the contributing scientist, and to quantify the contribution not only with respect to number of quotations, but also to the amount of data quoted.

7.2 Other ongoing projects & initiatives

Since data identification and citation is a hot topic, there are many studies in progress right now, investigating a multitude of different aspects of this field of data management. ENVRIplus partners in general, and WP6 in particular, should make sure that we 1) do not simply repeat what has already been done; 2) make the best use of the results and outcomes of other projects; and 3) collaborate with others where possible, while keeping the focus on issues that are of particular importance to (European) environmental infrastructures. An exciting opportunity for collaboration is the idea to work together with the THOR project in helping ENVRIplus partners integrate ORCID identifiers into their data workflows.

7.2.1 Project THOR

Project THOR (http://project-thor.eu/) is very relevant to WP6. A common denominator is the University of Bremen team, which is involved in both.

The Technical and Human Infrastructure for Open Research (THOR) project and its partners play a key role in ongoing efforts of integrating ORCID in ENVRIplus RIs. THOR is a 30-months H2020 project in its second year and aims at placing PIDs at the fingertips of researchers such that researchers have access to PIDs for their research artefacts relevant to scientific workflows. THOR has facilitated ORCID integrations in disciplinary repositories at EBI, CERN, and PANGAEA.

The THOR partners PANGAEA, ORCID, and DataCite are key to ongoing efforts towards coordinating, and facilitating ORCID integrations in RIs. A valuable and informative test case would be to help a selection of ENVRIplus RIs with their ORCID integration.

A first teleconference between ORCID, ICOS Carbon Portal, and PANGAEA was organised on October 20. The meeting was a first attempt to bring representatives of ICOS, as an ENVRIplus RI, and THOR together to discuss the idea of THOR contributions to ORCID integrations in RIs. There was strong support for such collaboration and a shared understanding of its benefits to both THOR and ENVRIplus, and its RIs.

Therefore, a short survey, submitted to 12 RIs, was undertaken in October 2016 with the aim to assess the relevance, timeliness, and state of ORCID integrations across ENVRIplus partner RIs. Among the 10 replies, 50% of the RIs plan to integrate ORCID, one of which is partially in implementation phase. This preliminary survey highlights that ORCID integrations are relevant and timely. Hence, we plan to build on the experiences gained in integrating ORCID in PANGAEA and continue discussions and coordination to increase understanding of the requirements and challenges faced by RIs in integrating ORCID. This effort is aimed at facilitating such integrations in ENVRIplus. This proposal was presented at the 3rd ENVRIweek in November 2016, and will be followed up by a workshop in early 2017.

7.2.2 Scholix – Scholarly Link Exchange

Scholix, Scholarly Link Exchange, (http://www.scholix.org/) was presented in June 2016 as a framework for the standardization of exchanged information about the links between research data and publications. The framework is developed by the working group Data Publishing Services coordinated by ICSU-WDS and RDA, in collaboration with actors such as Crossref, DataCite, PANGAEA, and International Association of STM Publishers [Burton 2017, Scholix 2016, STM Publishers 2017b] The framework is presented as a vision and guidelines, and builds on a multi-hub model with contributors including data centres, publishers and data repositories. The interoperability guidelines are based on the following assumptions:

- a number of natural hubs that aggregate data-literature and data-data links
- interoperability between those hubs and NOT between all the publishers of research objects
- the need for inference and retrospective linking (at least for several years to come)
- that some hubs will actively aggregate information from other hubs
- that other services will leverage this enabling infrastructure
- the need for unique persistent identification and standard referencing of research objects [Burton 2016, p. 6]

Services building on the Scholix framework include DataCite Event Data, Crossref EventData, and the OpenAIRE Data-Literature Interlinking (DLI) Service [Scholix 2016].

The original RDA/ICSU-WDS working group is no longer active, but has been succeeded by the RDA/WDS Scholarly Link Exchange (Scholix) Working Group, found at https://www.rd-alliance.org/groups/rdawds-scholarly-link-exchange-scholix-wg.

7.2.3 OpenAIRE

OpenAIRE is a project under the Horizon 2020 funding programme aimed at promoting open scholarship and support research data management for research conducted within Horizon 2020. OpenAIRE includes a repository of research objects (Zenodo) as well as workflow and interoperability services and guidelines, and a host of other issues related to open science [OpenAIRE 2017a]. Among the OpenAIRE work packages are Scholarly communication, which includes measuring Open Access impact and Literature-data integration. The leader of the literature-data integration activities is the European Bioinformatics Institute (EMBL-EBI), and other participants are Universität Bremen, Data Archiving and Networked Service (DANS), Consiglio Nazionale delle Richerche (CNR), and University of Athens [OpenAIRE 2017b]. In

January 5, 2017 there were 8966 datasets available discoverable through OpenAIRE's search facility ([OpenAire 2017d] https://www.OpenAIRE.eu/search/find?keyword=), as were document types such as reports, articles, data papers, software, images, annotations and patents, as well as projects [OpenAIRE 2017c].

To ensure OpenAIRE interoperability, repositories must be OAI-PMH compatible, and follow other OpenAIRE metadata requirements listed in the OpenAire Guidelines for Data Archives [OpenAIRE 2017d]. It can be noted that OpenAIRE allows for a multitude of PIDs: ARK, DOI, Handle, PURL, URN, and URL [OpenAIRE 2017e]. Data repositories need to be registered in Directory of Open Access Repositories, OpenDOAR (http://www.opendoar.org/) in order to be made available through OpenAIRE search [OpenAIRE 2017f]. There is a service for testing repository compatibility at https://www.openaire.eu/validator/welcome.

8 A SUGGESTED SYSTEM DESIGN FOR ENVRIPLUS

This Chapter presents a suggested system design for how ENVRIplus partners can apply and cite persistent identifiers for data. We have chosen to frame the system design in the form of "best practices" rather than a more formal service roadmap. However, we have included a brief summary of how Identification and Citation of data are described in the ENVRI reference model.

8.1 Introduction to our approach

Theme 2 of ENVRIplus is organised around five work packages that together cover six topics that are central to research data management: Cataloguing, Curation, Identification & Citation, Optimization, Processing, and Provenance. As shown in **Figure 7** below, these "pillars" are connected by three cross-cutting mechanisms: Architecture Design, Linking of Meta information, and a Reference Model.

In the case of Identification & Citation, there are especially strong links between this pillar and those of Curation, Cataloguing, and Provenance [Atkinson 2016]. The integration between these closely related topics will need to be well supported by tools, services and processing workflows. All these need to function together to accomplish the goals of the RIs, and ultimately support the research activities of end users.

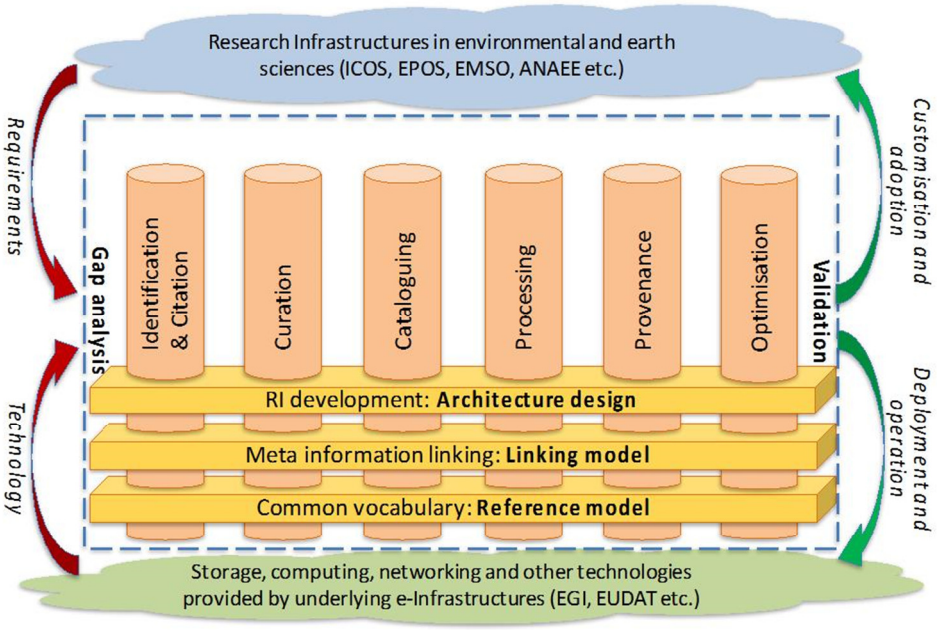

FIGURE 7. THE SIX PILLARS OF THEME 2 AND THE THREE CROSS CUTTING TOPICS. (ADAPTED FROM [ZHAO 2015].)

The ENVRI Reference Model [ENVRI RM V2.1 2016] offers a formal and strict way of describing roles, services and interconnections that occur throughout the data life cycle of a research-oriented organisation or project. In Chapter 8.2 below, we summarise how data identification and data citation are taken into account in the current version of the ENVRI RM, as outlined in Deliverable 5.2 [Hardisty 2016].

While being very useful from a more theoretical and abstract point of view, the mapping between the RM and a real-life research infrastructure rapidly becomes quite convoluted, and it may be difficult to clearly identify the correspondence between on the one hand actual RI practices and structures, and on the other hand the RM's concepts and their relationships.

As a complement and alternative to using the RM to provide a formal illustration of the data life cycle "architecture" of an RI, we have therefore chosen to focus on a more pragmatic and narrative approach. This centres on providing recommendations on best practices for (data) identification and citation that the ENVRIplus partners and other RIs in environmental and Earth sciences should undertake in order to make their data products findable, accessible, interoperable, and reusable. These best practices are listed in Chapter 8.3.

8.2 Data identification and citation in the ENVRI Reference Model

All research infrastructures for environmental sciences (the so-called 'ENVRIs') although very diverse, have some common characteristics, enabling them potentially to achieve a greater level of interoperability through the use of common standards and approaches for various functions. The objective of the ENVRI Reference Model (see, e.g., [ENVRI RM V2.1 2016]) is to develop a common framework and specification for the description and characterisation of computational and storage infrastructures. This framework can support the ENVRIs to achieve seamless interoperability between the heterogeneous resources of their different infrastructures. The ENVRI RM is structured according to the Open Distributed Processing (ODP) standard, ISO/IEC 10746-n, and as such, is defined from five different perspective or viewpoints: Science, Information, Computational, Engineering and Technology. Any of these viewpoints can be used to analyse and map out the five phases of the data lifecycle of any RI, i.e. Data Acquisition, Data Curation, Data Publishing, Data Processing and Data Use. (Note that as the RM focuses on data, the allocation of persistent identifiers to non-data objects, and how to refer to these in e.g. publications or other contexts, are not covered at present.)

In the RM, data identification and citation are related to mechanisms that help provide durable references to individual data objects, as well as collections of such objects. This allows proper acknowledgement of data creators and data publishing institutions. Thus, both data identification and data citation are recognised as core components of the common requirements of environmental RIs. Specifically, data identification falls under Data Curation (as requirement B.3) while data citation is part of Data Publishing (as requirement C.12) (see Appendix A of [ENVRI RM V2.1 2016]).

8.2.1 Data Identification in the RM[5]

The ENVRI RM describes data identification as "A functionality that assigns (global) unique identifiers to data contents".

The analysis of the ENVRIplus communities' requirements for data identification highlighted a need for identification management functionalities that include: DOI management; standard (homogenous) approach to Identification; identification of dynamic data series; identification of results from data queries (e.g., data services); data identification automation; identification of data objects stored as files (using file names as identifiers or suitable alternatives); identifier

[5] The material in this subchapter is based on chapter 6.6 in [Hardisty 2017].

systems used are based on handles (DOIs from DataCite, ePIC PIDs); persistent and unique identifiers for both data and metadata objects, and ensure availability of identification services.

Science Viewpoint: Data Identification is linked to the Data Curation phase, with two associated roles: PID Manager and PID Generator. The PID Manager is here a system or service that assigns persistent global unique identifiers to data and metadata products by invoking the external PID Service entity that is providing the actual PIDs. The PID Generator is a public system or service which generates and assigns persistent global unique identifiers to sets of digital objects, as well as maintains a public registry of these PIDs. In this context, Data Identification is a behaviour performed by a PID Manager which assigns a PID for data and metadata being curated.

Information Viewpoint: Unique Identifier (UID) is defined as an information object. Assigning a unique identifier is defined as an information action. An object may have more than one identifier associated with it.

Computational Viewpoint: PID service is defined as an external service to provide identifiers for data objects and to resolve objects. This service is used in brokered data import, processed data import, citation, as well as raw data connection, and should provide two interfaces: acquire identifier and resolve identifier. The PID service should be linked to the Data Curation phase. (It was originally linked to Data Use).

For provenance reasons, all derived data products must be identified. This includes both finalised products and intermediate ones. However, not all intermediate products are suitable for publishing. It is necessary to keep track of the data products but only those meant to be published require a public persistent identifier. Other data products (i.e., intermediate ones) may need a different type of identifier.

8.2.2 Data citation in the RM[6]

Requirements refer to the need for data citation functionalities including: citation management; standard (homogenous) approach to citation; data citation automation; guarantee unambiguous resolution of citations; ensure credit to curators and generators of derived data products; facilitate collection of usage statistics; facilitate citation of data subsets (coupling identification with query provenance).

The ENVRI RM version 2.0 describes data citation as “A functionality that assigns an accurate, consistent and standardised reference to a data object, which can be cited in scientific publications and/or from other data collections”.

Data citation involves producing a reference for a data source that can be resolved externally and link to the data within the RI, and is as such closely tied to data identification. The PID Manager is responsible for ensuring that the links to resources are kept consistent and informing of changes to the PID Generator. The RI can provide the template for citation of data; apart from this the citation activity is a responsibility of the data user.

Science Viewpoint: There is no role defined in the SV to handle citation. The Data Publishing community defines Data citation as a behaviour performed by a PID Generator, which is in charge of maintaining a reference between data object and identifier.

Information Viewpoint: Citation is defined as an information object. There is no information action linked to citation. There is no state linked to citation of data or metadata. The definition may need a state to indicate if the identifier has been used. In the modified version of the RM a Cite Data action is defined consuming provenance data.

Computational Viewpoint: Citation is mentioned as one of the ways for finding datasets in CV data publishing. However, this functionality mainly belongs at the data use phase, which is where citations are produced (to indicate that data has been used). PID Service is defined as an external

[6] The material in this subchapter is based on chapter 6.7 in [Hardisty 2017].

service to provide identifiers for data objects and to resolve objects. This service should provide two interfaces: acquire identifier and resolve identifier. The PID Service is linked to the data use phase. The PID Service is used in brokered data import, processed data import, citation, and raw data connection. This object is strongly related to citation resolution. Specific services for handling citations may be needed, especially for linking citations to provenance and resources.

8.3 Data Identification & Citation in practice — recommendations to RIs

Specifically, which type of persistent identifier is used by any RI should be dictated by the needs of both the RI and its typical end user communities. There are many different options (see Chapter 4.1). In general, those based on the Handle system (for example, DOIs from DataCite and PIDs from e.g. ePIC, as well as ORCIDs for people) are at present the most commonly used by ENVRIplus partners (based on the questionnaire). The amount of metadata that is mandatory to provide at the time of identifier registration ("minting") varies.

Persistent identifiers should be assigned throughout all parts of the research data life cycle. This supports all aspects of the FAIR (now often referred to as FAIR+R) principles (see **Chapter 2.4.2**), namely Findability, Accessibility, Interoperability, Reusability + Reproducibility — all required in the new era of truly data-intensive research that environmental research infrastructures are entering into. To comply with FAIR+R, it is clearly no longer adequate to rely on specific folder structure designs and/or file naming conventions.

In the analysis of the ENVRIplus partners' needs and requirements for Identification and Citation, [Atkinson 2016] point out that firstly, there is a need for ENVRIplus to design and implement a programme of awareness raising and practical training to alert those RIs that would benefit, and to raise the skills of practitioners in any RI, of the relevance of *Data Citation and Identification* issues and some of the available technologies that will help with solutions and rapid adoption of good practices.

Secondly, a key issue on the road towards a successful adoption of useful data identification and citation is the comprehensive adoption of appropriate working practices. Then, once good working practices are established, they should be formalised, e.g., as a workflow, and packaged, e.g., through good user interfaces, so that as much of the underpinning record keeping, e.g., *Citation*, *Cataloguing* and *Provenance* is automated. This has two positive effects, it enables the practitioners to focus on domain-specific issues without distracting record keeping chores, and it promotes a consistent solution that may be incrementally refined to adjust to changes in the RI's operations or the demands and needs of its designated user communities.

For all of the above mentioned outcomes to materialise, there have to be good technologies, services and tools supporting each part of these processes, e.g., data citations being automatically and correctly generated as suggested by [Buneman 2016]. Similarly, constructing immediate payoffs for practitioners using citation, as described in [Myers 2015], will increase the chances of researchers engaging with identification at an earlier stage.

In the following, we outline a set of best practices for Identification and Citation that can be adopted and followed as required by the ENVRIplus partners.

8.3.1 Identification best practices for RIs

ENVRIplus partners should **strive to implement** the use of **PIDs for all of the following** categories. (In some cases such as organisational entities, it may not yet be practical to assign PIDs, as the currently relevant registration schemes are poorly equipped to handle entities that frequently change names, stewardship etc.)

A. Data objects (files, databases etc.)
B. metadata objects

C. articles, reports and other documents related to the data
D. people, including everyone involved in the data production chain
E. organizations (agencies, institutes, and RIs themselves) involved in the data production chain
F. sensors and sensor platforms, measurement stations, cruises, measurement campaigns
G. physical samples

In addition, comprehensive use of **PIDs** should be **considered** for

H. queries used for accessing and retrieving (subsets of) datasets
I. data content types
J. software releases used in the data processing
K. workflows used in the data processing

To enable this approach, RIs need to take a number of steps, as outlined below for the different categories.

A. Data

The following applies to all kinds of digital data objects including flat files, file archives, and snapshots of databases (both relational and time-stamped/versionable types).

- set up, or subscribe to, a repository service for safe and trustworthy storage of the data objects;
- sign up to relevant PID providers (ePIC, DataCite, ...) themselves, or contact national data management resource providers (or data publishing repositories) to do this on their behalf
    - be prepared to pay yearly fees for prefixes
    - often also a small additional fee is paid per PID
    - costs & procedures vary a lot between countries!
- design workflows (scripts) or manual procedures for assigning PIDs — based on RESTful API calls
    - should include capability to provide RI-specific unique suffixes for each item
    - should include capability to calculate fixity information (checksums) and store in the PID registry
    - should include capability to update the data object's location information if it should change (due to a move to another repository, or a server name change)
- implement the PID assignment at all relevant data life cycle steps
- adapt existing cataloguing systems to store PID information
- set up a system (interfaced to the catalogue) that can dynamically create landing pages
    - landing pages should be interpretable both by humans and (via content negotiation [Wikipedia 2017a]) machine-operated processes and workflows
    - landing pages should both link on to the data object itself and provide all relevant metadata, or at least links to where the (descriptive) metadata can be retrieved

For data, practical issues like granularity and the data product level are important factors, as well as costs associated with assigning PIDs and maintaining their records. However, there are good reasons for applying PIDs at quite fine scales of granularity, as this is likely to greatly simplify and enhance

i. the overall data publication process, by allowing for correct citation of data accessed & used by researchers;
ii. the efficient retrieval of usage & citation statistics via bibliometrics indexing agents;
iii. the scientific reproducibility of the extraction of identical datasets based on a citation string including PID information;
iv. the correct usage of downloaded data, as all relevant metadata are reachable via the data object's landing page;
v. the annotation and associated record keeping of data items' provenance information;
vi. the assignment of credit to all involved personnel and institutes;
vii. the creation and maintenance of data collections;

viii. the application of machine-based data harvesting, processing and evaluation

B. Metadata

Metadata about data objects may be stored in several different ways, i.e. as separate documents, in a plain database or in a cataloguing system. Especially in the first case, it is good practice to handle the files as data objects in their own right. Of course, as periodic snapshots of the metadata and/or cataloguing system database(s) are backed up and stored in a trusted repository, they too, should be assigned a PID.

Follow the same steps as outlined above for data objects. Note that ideally the catalogue entry — and hence the corresponding landing page — for a metadata object should contain information about, or at least PID-based links to, all related data objects. In this way, the relationship between data and metadata is completely defined.

C. Articles, reports and other documents related to the data

All articles, reports and other documents related to the production of the data, the dataset types or the data from a project as a whole should be published — either in a data publication journal, or as separate documents on e.g. Zenodo, figshare or a similar repository site (such as the EUDAT B2SHARE service; see https://b2share.eudat.eu/). This will ensure that the documents will be assigned a persistent identifier, and thus be easily citable. This category also covers images and audio-visual content that describe the data production, analysis or general context of the data collection.

D. People, including everyone involved in the data production chain

- All personnel involved in the RI's data production activities should register at ORCID to receive their own individual identifier
- If possible and allowed by the RI's data publishing workflow, a coupling should be made between a dataset being published and assigned a DataCite DOI (by e.g. the RI itself or a data publisher/repository), and the ORCID IDs of the individuals listed as owner, producer, contributor or fulfilling any other role as defined in the DataCite DOI registry records. In this way, data publication credit can be automatically assigned.

E. Organizations (agencies, institutes, and RIs themselves) involved in the data production chain

Registering, and thus assigning a persistent identifier to, an organizational entity can be very complicated to achieve in practice, since some types of organizations and groups tend to be quite short-lived, or may be quite frequently reorganised.

- To facilitate unambiguous identification of the organizational entities (institutes, universities, agencies etc.) associated with the production of a dataset, it is recommended to register the organizations in ISNI.
- ORCID is also planning to offer support for organizations

F. Sensors & sensor platforms, and measurement campaigns & cruises

To simplify the unambiguous referencing to the specific conditions of data collection, for example in scientific literature and in the metadata records of datasets, it is recommended to

- set up and maintain a RI-specific catalogue of all sensors that are used to collect data, and subsequently assign a PID to each sensor (pointing to a landing page "driven" by the catalogue)
- assign PIDs also to platforms used for sensor deployment (buoys, aircraft, towers etc.)
- create separate PID records for time-limited measurement campaigns and cruises
- use PIDs as cross-referencing links to track of the relationships between sensors, platforms, campaigns etc.

G. Physical samples

Registering samples in a catalogue, and assigning unique identifiers to them, provides a simple yet very effective way to refer to the samples in publications, analysis workflows and related provenance records.

- Geological samples (and similar) can be registered at IGSN
- Biological samples, especially those from life science research, may be indexed using recommendations from BRIF (Bioresource Research Impact Factor initiative) [Mabile 2011].
- For other sample categories, a PID pointing to the sample's metadata record in a catalogue can easily be created.

H. Queries used for accessing and retrieving (subsets of) datasets

This follows the RDA Data Citation Working Group's recent recommendations.

- Format the query in a referable manner — for example a text file, or an executable URL
- Assign a PID to the query object.

I. Data content types

This refers to the application and implementation of data type registries (DTRs). These contain definitions of data types, typically created following specific metadata schemata in order to make them machine-operable. Data type definitions, which should always be presentable in forms that are interpretable both by human and machine users, are intended to guide the proper usage of a dataset. Examples include basic variable definitions such as "Temperature. Meteorological variable measured according to WMO guidelines. Unit: Centigrade" or more specific ones such as "Sonic air temperature. Deduced from sonic anemometer measurements via the eddy covariance technique. Unit: Centigrade."

- Select or define an adequate schema for the data type definitions. This can be Dublin Core, Inspire (ISO 19115) or something else.
- Using a data type registry, enter the data type definition following the selected schema. The definition is stored, and assigned a PID.
- The type definition PID can then be used in a data object's descriptive metadata, outlining its content

J. Software releases used in the data processing

Following the discussion in [Jones 2016], RIs wishing to apply identifiers for software that they develop, we recommend:

- For Open Source software, consider to store the software at a repository that supports versioning as well as download statistics tracking (e.g. GitHub)
- While not ideal for software, the metadata schema of DataCite is useable for registering software. Especially the fields Creator, Title, ResourceTypeGeneral and Description should be filled out.
- It may be desirable, or even necessary, to assign individual identifiers to all four levels of a software package: the product as a whole, individual versions or releases, specific operating system variants, and specific deployed instances

K. Workflows used in the data processing

Scientific workflows are often defined and stored using specialised workflow engine systems, such as Kepler (https://kepler-project.org/) and Taverna (https://taverna.incubator.apache.org/). These systems often use idiosyncratic formats to store the workflow configurations and outputs, which can create difficulties for users wishing to browse, link and execute existing workflows. To optimise reusability and interoperability, [Garijo 2017] recommends using Linked Data Principles when publishing workflows. Because Linked Data is based on using URIs, it is straightforward to use registered PIDs for referencing both workflows themselves as well as their and associated

input and output**.** Archive packages containing bundles of related configuration files and scripts for workflows can be published using e.g. Zenodo.

8.3.2 Citation best practices for RIs

The ENVRIplus partners should strive to follow the following recommendations for data citation, based on the review of data citation best practices and recommendations from relevant organisations including [Fenner 2016a], [FORCE11 2014a], [Socha 2013]:

Technical aspects:

A. All datasets intended for citation have a globally unique PID that can be expressed as an unambiguous URL
B. A PID expressed as a URL resolves to a landing page for a dataset
C. The landing page of a dataset is both human-readable and machine-readable (and preferably machine-actionable) and contains the dataset's PID
D. PIDs for datasets support multiple levels of granularity (including fine-grained subsets as well as collections)
E. Datasets are described with rich metadata (to track provenance information and to create meaningful citations and (including the identifier of the dataset))
F. Metadata are accessible even if a dataset is no longer accessible
G. RIs provide a robust resolver and registry for resolving PIDs and for data discovery
H. Metadata protocols and standards are used, that ensure interoperability with related stakeholders, e.g. cataloguing and indexing services
I. Data are published with a clearly defined data usage license

Citation practices:

J. RIs actively promote data citation (to users, publishers and other stakeholders in their research community (e.g. by providing documentation and how-tos) and by providing common citation formats to users)
K. Citation methods are flexible to support each community while still ensuring interoperability across communities

The citation best practice for RIs are outlined below.

A. All datasets intended for citation have a globally unique PID that can be expressed as an unambiguous URL

Based on the current and evolving practices and technological requirements of each RI, the choice of PID systems may differ across ENVRIplus partners. It is important to choose a PID system that facilitates interoperability.

B. A PID expressed as a URL resolves to a landing page for a dataset

When resolved through the respective handle system, PIDs will resolve into a URI that points to a landing page that either produces a human-readable or machine-readable summary of the relevant metadata of the data object and a link to the data object itself.

C. The landing page of a dataset is both human-readable and machine-readable

As stated, a dataset that is intended for citation is given a PID, and a corresponding landing page providing information about the dataset. The landing pages should be human-readable and machine-readable, and preferably machine-actionable.

D. PIDs for datasets support multiple levels of granularity (including fine-grained subsets as well as collections)

RIs should support data citation on multiple levels of granularity that suit the characteristics of their data and the needs of the community. Examples of levels of granularity include data points, subsets, datasets, and collection of datasets. Different PID systems may be used for different levels of granularity, providing they are interoperable.

E. Datasets are described with rich metadata on the landing pages

Metadata should include provenance information as well as other curation metadata about the dataset following at a minimum the metadata scheme of the PID system. Provenance information includes a list of all persons who have contributed to the dataset. In cases where there is a long list of contributors, these can be listed on the landing page of each data set instead of in the citation. Links to documentation that provides more information about the dataset are encouraged (see curation deliverable). Note that documentation and other research objects also can be given PIDs and associated landing pages.

F. Metadata are accessible even if a dataset is no longer accessible

Datasets might be deleted or replaced with a later version, in which case a new PID should be minted for the new version of the dataset. Thus, the previous versions of the dataset keep their PIDs and landing pages, and the new version is seen as a completely new dataset. Linking between the landing pages of the previous versions and the current version is encouraged to provide provenance information and facilitate tracking of the dataset usage through its versions. Note that the PID should still be resolvable to a landing page of the type tomb stone when a dataset is no longer accessible – this provides important information even though the dataset is no longer available.

G. RIs provide a robust resolver and registry for resolving PIDs and for data discovery

Citations and PIDs should be as persistent as the objects they cite, or actually even more persistent (see point F above regarding tomb stones for no longer accessible datasets). RIs should provide a robust resolver and registry, which could be handled through mirrored servers or a distributed storage solution.

H. Metadata protocols and standards are used, that ensure interoperability with related stakeholders, e.g. cataloguing and indexing services

RIs should choose metadata standards and protocols that ensure interoperability across RIs and services provided by other stakeholders, such as cataloguing and indexing services. A landing page should be able to generate the metadata in all relevant controlled vocabularies (see Cataloguing deliverable).

I. Data are published with a clearly defined data usage license

Data usage license information should be readily available to provide prospective users, preferably in both human-readable and machine-readable form.

Citation practices:

J. RIs actively promote data citation

ENVRIplus partners actively promote data citation practices to potential data users and other stakeholders in their research community. This could include providing documentation and how-to guides to data citation, and by providing common formats to users. RIs can facilitate and promote proper citation by including pre-created citation text snippets on the landing pages of all data sets they are curating. Citations can be automatically generated in a desired format using e.g. the DataCite DOI Citation Formatter API, which supports over 5,000 citation styles.

K. Citation practices are flexible to support each community, while still ensuring interoperability across communities

This entails promoting and developing citation practices as well as choosing and implementing technical solutions that suit the specific RIs, while taking into account the ENVRIplus community as a whole. RIs should interact with their user base — in OAIS terminology, called the "designated community" — to investigate its data citation practices. In many cases, scientific end users are following old, pre-PID habits by routinely referring to data sets in the running texts of articles (e.g. "we used the ABC dataset provided by Andersson et al." in a paper's "Materials and

Methods" section). This behaviour should be actively discouraged, instead encouraging proper data citation practices by highlighting good examples of proper data citation using PIDs (DOIs) in the context of conference presentations or workshops., and by following the data citation best practices recommended here.

8.4 Technologies, services and tools

RIs need access to services that provide registration, indexing and discovery (through lookup of the associated handles) of objects — preferably at the moment the objects are ingested into a trusted repository.

In the case that a RI does not operate a sophisticated-enough cataloguing service, capable of supporting landing pages with the necessary level and detail of metadata and other information, then this capability needs to be provided by an external service provider. As an example, the metadata catalogue operated by DataCite supports a quite exhaustive set of attributes that may be used to populate landing pages for data objects that have been assigned a DOI.

To extract usage statistics, the RI needs to combine information on the frequency of access of a data resource with bibliometrics information based on citations and mentions in (scientific) literature and other media. The access frequency analysis may include everything from positive search results, visualization and downloads via the RI's own data centre portal (if applicable) or the corresponding statistics from RI-external repositories holding the data.

9 NEGOTIATIONS WITH PUBLISHERS AND OTHERS

Environmental research infrastructures, including the ENVRIplus partners, are often built on a large number of distributed observational or experimental sites, run by hundreds of scientists and technicians, financially supported and administrated by a large number of institutions. As data from these RIs are made available (and especially so if this is done under an open access policy), it becomes very important to properly acknowledge the data sources and their providers. At the same time, it will be crucial to implement common and efficient data citation tracking systems that allow data providers to identify downstream usage of their data so as to prove their importance and show the impact to stakeholders and the public.

Achieving these goals will require actions to be taken across the board:

- RIs need to adjust their own data management practices, as well as the behavioural patterns of their end user communities
- PID registration providers need to tailor their services and metadata schemata to match RIs' needs
- Scientific publishers should adapt their requirements for data linking as well as their practices for handling citations to data and other (digital) research objects
- Agencies that collect references and citations need to provide comprehensive bibliometrics and statistics that allow to distinguish data usage from traditional article citations
- Organisations that bring together professionals interested in research data management must be made aware of the needs and requirements of environmental RIs

To assist in the development of relevant applications and mechanisms, environmental RIs need to make their requirements and wishes known to all the relevant actors. It is therefore one of the important tasks for Work Package 6 to organise a suitable meeting platform, and then to conduct discussions and negotiations on collaboration and contracts.

9.1 The case for negotiations

The ENVRIplus partner research infrastructures are complex entities, receiving support and funding from a multitude of stakeholders and agencies, and typically involving a number of different institutions and hundreds of people. When data from these RIs are shared under an

open access policy, it becomes very important to make sure that end users acknowledge the data sources and their providers, so that appropriate credit can be assigned. To allow data providers to identify downstream usage of their data, and hence to assess the impacts their data have, common data citation tracking systems are needed to accumulate usage statistics that can be shared with stakeholders and the public.

Despite a current reflux and attempts to change the publish-or-perish culture in science by removing the emphasis on numeric impact analysis techniques that lead to a large influence of a scientist's h-index or similar on her or his career opportunities, citation indexing is still an important and dominant means to determine the 'quality' of scientists and their work. Around this a big industry has evolved where publishers and their related indexing services have found a way to control the scientific community and their stakeholders to bind them to their paid services like Web of Science and Scopus.

One way to ensure that scientists and other contributors receive better credit for the outcomes of their work that are complementary to their publications, would be to provide a new kind of bibliometrics index related to citations of data products in scholarly publications — something like a d-index. This is currently not provided or supplied by the generic indexing services, as the DOIs that refer to data in citation are not counted in the same way as citations of (peer-reviewed) publications. There are also some complexities in calculating a 'd-index' that have to do with the specific properties of data, the fact that data are normally not peer-reviewed, and that there is no ranking of the quality of the publishing platform like the SI or h5 index for scientific journals.

DataCite is a non-profit organisation that provides the services that allows the minting, identification and publishing of DOIs for data objects and their metadata. In this way "citable data become legitimate contributions to scholarly communication, paving the way for new metrics and publication models that recognise and reward data sharing" [http://www.datacite.org].

It will require an joint effort from data providers, services like DataCite, stakeholders, data centres, publishers and citation service providers to setup an indexing service for scientific data and to have this accepted by the community of scientists as the right platform. ENVRIplus represents an important group of data providers that could play an important avant-garde role in showcasing such a development.

9.2 Discussion partners

The first task of the "negotiation" process is to identify the relevant discussion partners, which should include:

- PID providers
- Publishers & publisher associations
- Data usage indexers
- Library organizations
- Research Data Alliance, CODATA and similar organisations

Contacts will therefore be made with all the organisations that were identified in our review of the service provider and publisher "landscape", see **Chapter 6**.

9.3 Wish-list for services

As indicated by the collection of Requirements reported in [Atkinson 2016], as well as conversations and discussions during the ENVRIplus collaboration meetings, the project partners have quite varying needs for tools, services and guidance related to identification of data and support for citation of datasets. However, a number of common wishes and suggestions have been identified, as outlined below.

Still, before embarking on discussions and negotiations with external partners, we suggest to get back in contact with all ENVRIplus RIs in order to find out if their understanding and insights into data identification & citation have significantly changed over the last year, and/or whether their needs for services and support are now different — as a result of either a changed prioritisation of I&C-related issues, or changes in the overall focus of the RIs' scientific & societal missions. In addition, it will be crucial to the continued Work Package 6 work to identify and maintain an up to date list of all Identification & Citation experts in the ENVRIplus partner organisations. This will streamline the information gathering step and ensure that the information collected is pertinent and correct, as well as set up an efficient communication channels between all ENVRIplus RIs and WP6.

Based on the outcome of this new survey, it is to be expected that updates and reprioritisations will have to be made to the initial wish list of services and functionalities outlined below.

9.3.1 Generic identifier minting services serving individual RIs

For organizations like RIs who have decided to procure and register their own Handle prefixes (ICOS is an example), it is highly desirable to be able to assign "their own" PIDs to both data and documents. Interestingly, even though most reading material clearly states that DataCite DOIs are to be used for research data, at the same time DataCite's Metadata Kernel version 3.1 only specifies that the resource described by the metadata kernel "can be of any kind, but it is typically a dataset. We use the term 'dataset' in its broadest sense." [DataCite 2016e, p. 4] In any case, the majority of DataCite DOIs currently specify that the registered objects are of type 'text', and many of these are reports or other free-text documents rather than e.g. tabular data in comma- or tab-separated asci text format.

9.3.2 Dynamic data, including support for versioning

Providers of PIDs for data should accept registration of query strings (to repository data bases) as "proxies" for data objects. Similarly, it should be investigated if such queries fulfil the obligations from journal publishers for authors to make available research data used in scholarly publications. The outcomes of the RDA Data Citation working group [Rauber 2016] will provide a good starting point.

In parallel, the complex issue of how to best refer to dynamic data should be explored further, with the goal to simplify access and retrieval of both the latest version of a specific dataset as well as earlier versions. As an example, a standard method for including metadata such as "this object is replaced by" and "this object replaces" at the PID registry level is desirable (as already implemented by e.g. DataCite). Tools that facilitate the tracing of the resulting "data trails" should be implemented to support e.g. workflow engines.

9.3.3 Support for inclusion of sub-setting information in citations

Publishers should allow authors to append "sub-setting" information to identifiers used in citations, for example by adding selection parameters at the end of URLs, e.g. in the form of 10.1234/zenodo.1234?from=X&to=Y… PID lookup services used to resolve the citation URL should then be able to pass on the parameters to the resource location (landing page) in the cited object's PID record. Finally, the data centre or repository hosting the landing page should be prepared to act on the selection parameters in order to serve an end user with the data corresponding to the citation.

This is a topic that has been under active discussion for several years, see e.g. the report from the joint COOPEUS/ENVRI/EUDAT PID workshop held in Bremen 25-26 June, 2013 [Huber 2013]. Despite this, there is still no complete agreement on standardised approaches for how to achieve this, although there are many successful examples; see, e.g., the example of marine data from the Argo project in **Chapter 7.1.3**. However, there is considerable resistance from some "Handle purists", who argue that PIDs should not carry any semantic information, neither in the prefix and suffix combination itself, nor in the form of appended strings.

9.3.4 Management of data collections

Collections offer great advantages but can be problematic from a bibliometrics point of view. Services should be developed that allow both data producers and end users of data to identify all registered collection objects that contain a specific individual data object (based on its PID). Other services that can recursively extract "complete author lists" for a collection by summarising the creator information for all its individual member objects are also needed, in order to support efficient extraction of usage statistics for data objects. Possible solutions for how to better support data collection registration are being developed within the e.g. the ENVRIplus Implementation Case IC-09 (see **Chapter 7.1.4**) and the RDA Research Data Collection working group (https://www.rd-alliance.org/groups/pid-collections-wg.html).

9.3.5 Sustainable data typing services

If data types are to become adopted by a wide range of RIs and other research organizations, there needs to be a structure in place that will guarantee the sustainability of the registries that host the type definitions.

Data typing offer great advantages but are problematic from several points of view. Firstly, there needs to be some consensus on the standards used for the definitions themselves, i.e. the schemata used. Secondly, the operation and maintenance of the registries must be guaranteed in a long-term perspective. Thirdly, there is disagreement over the relative advantages/disadvantages, sustainability/long-term stability and complementarity between definition systems based either on semantic web ontologies/linked open data or data type registries.

9.4 The negotiation timeline

As specified by the WP6 Description of Work, a report on the negotiations with PID service providers, publishers, providers of existing data citation systems and other scientific organisations (D6.2) is due at the end of M36. We therefore plan to organise the related work as follows:

- Revisit the RI data identification & citation requirements survey — February-April 2017
- Workshop with publishers, PID providers, indexers and ENVRIplus RIs — May 2017
- Shortlist of negotiation partners, and finalised strategy ready — June 1, 2017
- First rounds of negotiations — June-September 2017
- Report milestone — October 31, 2017
- Second rounds of negotiations — November 2017-January 2018
- Writing of Deliverable D6.2 — February-March 2018
- Deliverable D6.2 — submission April 30, 2018

10 CONCLUSIONS

In this Deliverable, we present both a suggested system design for how ENVRIplus partners can register and cite persistent identifiers for data, and a strategy for upcoming negotiations and discussions with PID service providers, scientific publishers and bibliometrics analysts. We also provide an introduction into the topics of data identification and data citation, map the landscape of service providers, publishers and indexing agents, and highlight related work that is being, or has been, undertaken outside of ENVRIplus as well as by ENVRIplus partners.

We have chosen to frame the system design in the form of "best practices" rather than a more formal service roadmap or a detailed comparison with the ENVRI reference model. The proposed best practices are instead based on a combination of work performed by international expert groups (such as the Research Data Alliance, FORCE11, and CODATA) and the outcomes of the study performed by Work Package 5 of ENVRIplus partner data identification & citation-relevant requirements and related technologies.

The negotiation strategy comprises a set of topics and questions that we identified (again based on the recent requirements study) as high priority for ENVRIplus partners, and a time plan for carrying out the necessary steps and negotiation components within the scope of planned Work Package 6 activities.

Finally, this document is on purpose rather extensive. The reasons for this are manifold, but first and foremost we intend it to be able to serve as a starting point for researchers who are new to the concepts of Data Identification and Citation. Secondly, the Deliverable aims to be a source of information, both about basic concepts and the landscape of service providers, citation indexers and research data organisations concerned with developing the PID concept.

11 IMPACT ON PROJECT

Attribution of data produced by Research Infrastructures is vital for the data providers serviced by the RIs and the RIs themselves, as this enables them to inform their sponsors and stakeholders about the impact of their work. Tracking the use by citations of the datasets and the number of scientific papers based on the data is an important task for many RIs. Facilitating this is a difficult task for which coordinated developments in both the information technology implementation of the RIs and in the workflows and practices of citation data providers are needed. This deliverable explores the requirements and presents a plan that will help and guide the RIs to achieve the much needed developments and integrate this into the data lifecycle of their organisations.

The impact of this service to the project is potentially extremely large and is actually at the core of the raison-d'être of RIs who should provide proper identification of all their data as the basis for curation, provenance, processing, publishing and data usage tracking services, that go far beyond number of downloads and downloaded data volume. The service should connect to checking of data licenses, authentication and authorisation of users requesting access and provide suggested citation of the (dynamic) data. When an RI also offers dynamic generation of subsets and/or collections of data and correct citations and attribution of this data, this will require a whole suite of additional technological developments that are described in the RDA recommendations on dynamic data [Rauber 2016], such as a fully versioned metadata store and minting PIDs for persistent data queries.

12 IMPACT ON STAKEHOLDERS

The topic of Data Identification and Citation does not stand alone, but is strongly and inextricably linked to the other Theme 2 topics, and especially so to Cataloguing, Curation and Provenance. These connections have important implications for all ENVRIplus RIs, not only for their own internal data management, but also for their ability to exchange data and services with the other project partners and beyond.

With this Deliverable, we want to encourage the ENVRIplus partners to engage in a dialogue and exchange of ideas, experiences and practical implementations related to identification and citation. There are many benefits and advantages from sharing working practices and technology in this area across RIs; for example a) to reduce maintenance and development costs; b) to facilitate interdisciplinary research that draws on multiple RIs; c) to improve "weight" when negotiating; and d) to help nurture a culture of fair attribution and reproducible science.

The Work Package 6 team stands ready to facilitate this communication and collaboration, for example by helping to connect PID-interested people across the ENVRIplus partners. We are of course also, through our participating RI experts, ready to provide advice on services and practices, and we invite all interested parties to join our implementation cases IC-01 and IC-09.

We close with these quotes from two project partner RIs, summarising their expectations:

"For ICOS, Deliverable 6.1 presents the basis for using PIDs in the data infrastructure and using this as the basis for a flexible data tracking and citation approach based on the RDA recommendations for dynamic data citation." -- Alex Vermeulen (ICOS)

"For DKRZ, Deliverable 6.1 is a good overview on what can be done with PID as a best practice. We will carefully go through it, although we already have implemented various aspects of the topics. For (wo)manpower reasons, we cannot yet say what we will be able to implement. However, this can be a good guidance for any further planning." -- Frank Toussaint (IS-ENES)

REFERENCES

[Almas 2015] B. Almas, J. Bicarregui, A. Blatecky, S. Hill, L. Lannom, R. Pennington, R. Stotzka, A. Treloar, R. Wilkinson, P. Wittenburg and Z. Yunqiang: "Data Management Trends, Principles and Components – What Needs to be Done Next?" Report from the Research Data Alliance Data Fabric Interest Group, draft version (paris-doc-v6-1_0.docx) from September 2015. Available via http://hdl.handle.net/11304/f638f422-f619-11e4-ac7e-860aa0063d1f.

[ALPSP 2017a] Association of Learned & Professional Society Publishers: About ALPSP Membership. http://www.alpsp.org/Membership. Accessed 2017-01-04.

[Amirtha 2015] Amirtha, T. 2015. The open publishing revolution, now behind a billion-dollar paywall. Fast Company magazine. Published 2015-04-17. https://www.fastcompany.com/3042443/mendeley-elsevier-and-the-future-of-scholarly-publishing. Accessed 2017-01-30.

[Atkinson 2016] M. Atkinson, A. Hardisty, R. Filgueira, C. Alexandru, A. Vermeulen, K. Jeffery, T. Loubrieu, L. Candela, B. Magagna, P. Martin, Y. Chen and M. Hellström: A consistent characterisation of existing and planned RIs. ENVRIplus Deliverable 5.1, submitted on April 30, 2016. Available at http://www.envriplus.eu/wp-content/uploads/2016/06/A-consistent-characterisation-of-RIs.pdf Accessed 2017-01-30.

[Austin 2016] Austin, C. C., Bloom, T., Dallmeier-Tiessen, S., Khodiyar, V., Murphy, F., Nurnberger, A., Whyte, A. (2015). Key components of data publishing: Using current best practices to develop a reference model for data publishing. DOI: https://doi.org/10.5281/zenodo.34542.

[Bechhofer 2013] Bechhofer, S, Buchan, I, De Roure, D, Missier, P et al. 2013. Why linked data is not enough for scientists. *Future Generation Computer Systems*, Volume 29, Issue 2, February 2013, Pages 599-611, ISSN 0167-739X, http://dx.doi.org/10.1016/j.future.2011.08.004.

[Beck 2016] K. Beck, R. Ritz and P. Wittenburg: Towards a Global Digital Object Cloud – Report from the Views on PID Systems training course and workshop. Available at https://www.rd-alliance.org/sites/default/files/attachment/20160901_RDA_PID_event_Garching_report_final.pdf

[Berners-Lee 1998]. T. Berners-Lee: Cool URIs don't change, World Wide Web Consortium (W3C), Cambridge, MA. Available at https://www.w3.org/Provider/Style/URI.html Accessed 2017-01-30.

[Borgman 2007] Borgman, C. 2007. *Scholarship in the digital age*. Cambridge, MA: MIT Press.

[Borgman 2015] Borgman, C. 2015. *Big data, little data, no data: scholarship in the networked world*. Cambridge, MA.: MIT Press.

[Buneman 2016] Buneman, P.; S. Davidson and J. Frew, Why data citation is a computational problem, Communications of the ACM, Vol. 59 No. 9, Pages 50-57, http://dx.doi.org/10.1145/2893181.

[Burton 2016] Burton, A & Koers, H. 2016. Interoperability Framework Recommendations. ICSU-WDS & RDA Publishing Data Services WG. http://www.scholix.org/guidelines Accessed 2017-01-04.

[Burton 2017] Burton, A., Koers, H., Manghi, P., Stocker, M., Fenner, M., Aryani, A., La Bruzzo, S., Diepenbroek, M., Schindler, U. and Authr, C., 2017. The Scholix Framework for Interoperability in Data-Literature Information Exchange. *D-Lib Magazine*, *23*(1/2). https://doi.org/10.1045/january2017-burton

[Bütikofer 2009] N. Bütikofer: Catalogue of criteria for assessing the trustworthiness of PI systems, nestor-Materialien, Niedersächsische Staats und Universitätsbibliothek Göttingen, Göttingen, Germany. http://nbn-resolving.de/urn:nbn:de:0008-20080710227 Accessed 2017-01-30.

[Callaghan 2013] Callaghan, S., Murphy, F., Tedds, J., Allan, R., Kunze, J., Lawrence, R., Mayernik, M.S., Whyte, A. 2013. Processes and procedures for data publication: a case study in the geosciences. *Int. Journal of Digital Curation*, Volume 8, Issue 1. http://dx.doi.org/10.2218/ijdc.v8i1.253

[Creative Commons 2016a] Creative Commons. 2016. Licensing types – Creative Commons. https://creativecommons.org/share-your-work/licensing-types-examples/ Accessed 2016-12-20.

[Creative Commons 2016b] Creative Commons. 2016. CC0 – Creative Commons. https://creativecommons.org/share-your-work/public-domain/cc0/ Accessed 2016-12-20.

[Creative Commons 2016c] Creative Commons. 2016. Creative Commons Attribution 3.0 Unported (CC BY 3.0). https://creativecommons.org/licenses/by/3.0/ Accessed 2016-12-20.

[Creative Commons 2016d] Creative Commons. 2016. Creative Commons Attribution-ShareAlike 3.0 Unported (CC BY-SA 3.0. https://creativecommons.org/licenses/by-sa/3.0/ Accessed 2016-12-20.

[CODATA 2016a] Committee on Data for Science and Technology. 2016. Our mission. http://www.codata.org/about-codata/our-mission Accessed 2016-12-19.

[CODATA 2016b] Committee on Data for Science and Technology. 2016. CODATA Becomes a Participating Organisation of GEO. http://www.codata.org/news/24/62/CODATA-Becomes-a-Participating-Organisation-of-GEO Accessed 2016-12-19.

[Copernicus 2017] Copernicus Publications. Copernicus Publications – Journals by subject. http://publications.copernicus.org/open-access_journals/journals_by_subject.html Accessed 2017-01-04.

[Costas 2015] Costas, R., Zahedi, Z., & Wouters, P. 2015. Do altmetrics correlate with citations? Extensive comparison of altmetric indicators with citations from a multidisciplinary perspective. *Journal of the Association for Information Science and Technology*, 66(10), 2003-2019. https://dx.doi.org/10.1002/asi.23309

[Cronin 2005] Cronin, B. 2005. *The hand of science: academic writing and its rewards*. Lanham, Md: Scarecrow Press.

[Crossref 2013] Crossref. 2013. crossref.org :: Fast Facts. http://www.crossref.org/01company/16fastfacts.html Accessed 2017-01-02.

[Crossref 2015a] Crossref. 2015. crossref.org ::history/mission. http://www.crossref.org/01company/02history.html Accessed 2017-01-02.

[Crossref 2015b] Crossref. 2015. crossref.org ::publishers & societies. http://www.crossref.org/01company/06publishers.html Accessed 2017-01-02.

[Crossref 2015c] Crossref. 2015. crossref.org :: libraries. http://www.crossref.org/01company/07libraries.html Accessed 2017-01-02.

[DataCite 2016a] DataCite. 2016. Our mission. https://www.datacite.org/mission.html Accessed 2016-12-19.

[DataCite 2016b] DataCite. 2016. Members. https://www.datacite.org/members.html Accessed 2016-12-19.

[DataCite 2016c] DataCite. 2016. Citation formatter. https://www.datacite.org/citation.html Accessed 2016-12-19.

[DataCite 2016d] DataCite. 2016. OAI-PMH Data Provider. http://oai.datacite.org/ Accessed 2016-12-19.

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

[DataCite 2017] DataCite. 2017. DataCite Event Data. https://eventdata.datacite.org/docs/deposits Accessed 2017-01-13.

[DataCite Profiles 2016] DataCite Profiles. 2016. https://profiles.datacite.org/ Accessed 2016-12-19.

[Dataverse 2016a] Dataverse. 2016. About | The Dataverse Project – Dataverse.org. http://dataverse.org/about Accessed 2016-12-20.

[Dataverse 2016b] Dataverse. 2016. Researchers | The Dataverse Project – Dataverse.org. http://dataverse.org/researchers Accessed 2016-12-20.

[Dataverse 2016c] Dataverse. 2016. Academic Credit | The Dataverse Project – Dataverse.org. http://dataverse.org/best-practices/academic-credit Accessed 2016-12-20.

[Dataverse 2016d] Dataverse. 2016. Configuration http://guides.dataverse.org/en/latest/installation/config.html#doiprovider Accessed 2016-12-20.

[Dataverse 2016e] Dataverse. 2016. Data Citation – Dataverse.org. http://best-practices.dataverse.org/data-citation Accessed 2016-12-20.

[DKRZ 2016] Deutsches Klimarechenzentrum. 2016. Services around climate research. https://www.dkrz.de/about-en/dienste Accessed 2016-12-19.

[Dodds 2014] L. Dodds, G. Phillips, T. Hapuarachchi, B. Bailey and A. Fletcher, "Creating Value with Identifiers in an Open Data World". Report from Open Data Institute and Thomson Reuters, October 2014. Available at http://innovation.thomsonreuters.com/content/dam/openweb/documents/pdf/corporate/Reports/creating-value-with-identifiers-in-an-open-data-world.pdf

[DORA 2012] San Francisco Declaration on Research Assessment. 2012. http://www.ascb.org/files/SFDeclarationFINAL.pdf?x30490 Accessed 2017-01-13.

[Duerr 2011] R.E. Duerr, R.R. Downs, C. Tilmes, B. Barkstrom, W.C. Lenhardt, J. Glassy, L.E. Bermudez and P. Slaughter, "On the utility of identification schemes for digital earth science data: an assessment and recommendations". *Earth Science Informatics*, vol 4, 2011, 139-160. https://dx.doi.org/10.1007/s12145-011-0083-6

[ENVRI RM V2.1 2016] ENVRI Reference Model V2.1, November 9 2016. https://wiki.envri.eu/download/attachments/8553250/EC-091116-1403.pdf Accessed 2017-01-10. Also available in wiki format at https://wiki.envri.eu/display/EC/ENVRI+Reference+Model.

[ePIC 2016a] European Persistent Identifier Consortium. 2016. ePIC structure. http://www.pidconsortium.eu/?page_id=74 Accessed 2016-12-19.

[ePIC 2016b] European Persistent Identifier Consortium. 2016. Service. http://www.pidconsortium.eu/?page_id=88 Accessed 2016-12-19.

[Fenner 2016a] Fenner, M, Crosas, M, Grethe, JS, Kennedy, D, Hermjakob, H, Rocca-Serra, P, Berjon, R, Karcher, S, Martone, M & Clark, T. (2016). A data citation roadmap for scholarly data repositories. bioRxiv preprint published 2016-12-28. https://doi.org/10.1101/097196

[Fenner 2016b] Fenner, Martin: Announcing the Organization Identifier Project: a Way Forward. Blog published 2016-11-01. https://blog.datacite.org/announcing-organization-identifier-project/ Accessed 2016-12-19

[figshare 2017a] Figshare. 2017. Figshare – About. https://figshare.com/about Accessed 2017-01-02.

[figshare 2017b] Figshare. 2017. How persistent is my research? : fighshare Support. https://support.figshare.com/support/solutions/articles/6000079089-how-persistent-is-my-research- Accessed 2017-01-02.

[figshare 2017c] Figshare. 2017. Is the research I put on Figshare citable / Do I get a DOI? Figshare Support. https://support.figshare.com/support/solutions/articles/6000079036-is-the-research-i-put-on-figshare-citable-do-i-get-a-doi- Accessed 2017-01-02.

[FORCE11 2014a] Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. Martone M. (ed.) San Diego CA: FORCE11; 2014. https://www.force11.org/group/joint-declaration-data-citation-principles-final. Accessed 2016-12-30.

[FORCE11 2014b] FORCE11. 2014. Guiding Principles for Findable, Accessible, Interoperable and Re-Usable Data Publishing Version B1.0. https://www.force11.org/fairprinciples Accessed 2016-11-28.

[FORCE11 2016] FORCE11. 2016. Data Citations: A Primer. http://force11.github.io/data-citation-primer/ Accessed 2016-12-30.

[FOSTER 2016] FOSTER. 2016.Open Science taxonomy tree. http://openscience.com/wp-content/uploads/2016/10/xfLOCi.png Accessed 2016-11-28.

[Gallagher 2015] J. Gallagher, J. Orcutt, P. Simpson, D. Wright, J. Pearlman and L. Raymond, "Facilitating open exchange of data and information". Earth Science Informatics, Volume 8, Issue 4, pp 721-739, December 2015. Available via http://dx.doi.org/10.1007/s12145-014-0202-2.

[Garijo 2017] D. Garijo, Y. Gil and O. Corecho: Abstract, Link, Publish, Exploit: An End to End Framework for Workflow Sharing. To appear in *Future Generation Computer Systems*. Available on request from dgarijo@isi.edu.

[Haak 2016] Haak, Laure: Organization identifier project: A way forward. Blog published 2016-10-31. http://ORCID.org/blog/2016/10/31/organization-identifier-project-way-forward Accessed 2016-12-19

[Hahnel 2012] Hanhel, M. 2012. Ensuring persistence on figshare. Blog post published 2012-04-13 https://figshare.com/blog/Ensuring%20persistence%20on%20figshare/25 Accessed 2017-01-02.

[Hahnel 2014] Hahnel, M. 2014. Working with GitHub and Mozilla to enable 'Code as a research output'. Blog post published 2014-03-13 https://figshare.com/blog/Working_with_Github_and_Mozilla_to_enable_Code_as_a_Research_Output_/117 Accessed 2017-01-02.

[Hahnel 2015] Hahnel, M. 2015. Going global with DataCite. Blog post published 2015-03-18. https://figshare.com/blog/Going_global_with_DataCite/150 Accessed 2017-01-02.

[Halperin 2016] Halperin, J. R. and Prywes, N. 2016. Making data and tools available for the world to see: Arturo Sanchez of CERN on why ATLAS uses CC0. Data https://creativecommons.org/2016/11/02/atlas-cern/ Blog post published 2016-11-02. Accessed 2016-12-20.

[Hardisty 2017] A. Hardisty, A. Nieva de la Hidalga, D. Lear, B. Magagna, M. Atkinson and K. G. Jeffery: Reference-model guided RI design. ENVRIplus deliverable D5.2, submitted on January 5, 2017.

[Huber 2013] R. Huber, A. Asmi, J. Buck, J.M. de Luca, D. Diepenbroek, A. Michelini, and participants of the Bremen PID workshop, "Data citation and digital identification for time series data & environmental research infrastructures", report from a joint COPEUS-ENVRI-EUDAT workshop in Bremen, June 25-26, 2013. Available via http://dx.doi.org/10.6084/m9.figshare.1285728

[Hyan 2015] R. Hyan: Taxa, taxon names and globally unique identifiers in perspective. In: Watson et. al (Eds), *Descriptive Taxonomy: The Foundation of Biodiversity Research*, Cambridge University Press. p.260-270, http://doi.dx.org/10.13140/2.1.3381.8400.

[ISNI 2016a] International Standard Name Identifier. 2016. ISNI. http://www.isni.org/ Accessed 2016-12-19.

[ISNI 2016b] International Standard Name Identifier. 2016. About the ISNI International Agency. http://www.isni.org/about Accessed 2016-12-19.

[Jessop 2016] P. Jessop of County Analytics Ltd.: personal communication (to M. Lassi) in connection to the PIDapalooza meeting in Reykjavik, Iceland, November 9-10, 2016.

[Jones 2016] C. Jones, B. Matthews, I. Gent, T. Griffin and J. Tedds: Persistent Identification and Citation of Software. Paper presented at the 11th International Digital Curation Conference in Amsterdam, The Netherlands, 22-25 February, 2015. PURL: http://purl.org/net/epubs/work/24496942.

[Kahn 1995] R. Kahn and R. Wilensky: A Framework for Distributed Digital Object Services, Technical Note, Corporation for National Research Initiatives, Reston, VA. Re-published in *International Journal on Digital Libraries* (2006) 6(2): 115–123, http://dx.doi.org/10.1007/s00799-005-0128-x

[Koshoffer 2016] Koshoffer, A. 2014. ResearchGate is now generating DOIs for content. Blog published 2014-11-03. https://libapps.libraries.uc.edu/blogs/dlc/2014/11/03/researchgate-is-now-generating-dois-for-content/ Accessed 2016-12-20.

[Kratz 2015] Kratz, J. E. & Strasser, C. 2015. Making data count. *Scientific Data*, article number 150039. http://doi.org/10.1038/sdata.2015.39

[Kunze, J. (2003). Towards electronic persistence using ARK identifiers. In: *Proceedings of the 3rd ECDL Workshop on Web Archives*, Trondheim, Norway, August 21, 2003. Available at http://bibnum.bnf.fr/ecdl/2003/proceedings.php?f=kunze. Accessed 2017-01-30.

[Lagotto 2017] Lagotto. 2017. About Lagotto. http://www.lagotto.io/ Accessed 2017-01-13.

[Lannom 2016] L. Lannom and P. Wittenburg: Global Digital Object Cloud (DOC) — A guiding vision. http://dx.doi.org/ 11304/33251dbb-07d7-4838-97cf-d3275067a35a.

[Lin 2013] Lin, J. & Fenner, M. 2013. Altmetrics in Evolution: Defining & Redefining the Ontology of Article-Level Metrics. *Information Standards Quarterly*, Summer 2013, 25(2): 20-26. http://dx.doi.org/10.3789/isqv25no2.2013.04

[Mabile 2011] L. Mabile et al.: The BRIF (Bioresource Research Impact Factor) as a tool for improving bioresource sharing in biomedical research. *Nature Precedings* : hdl:10101/npre.2011.6568.1. Available via https://core.ac.uk/download/pdf/290429.pdf Accessed 2017-01-03.

[MakingDataCount 2015a] Making Data Count. 2015. Making Data Count – Home page for the Data Level Metrics Project. http://mdc.lagotto.io/ Accessed 2017-01-13.

[Mendeley 2017a] Mendeley. 2017. Reference Manager | Mendeley. https://www.mendeley.com/reference-management/reference-manager/ Accessed 2017-01-02.

[Mendeley 2017b] Mendeley. 2017. Datasets | Mendeley. https://www.mendeley.com/datasets Accessed 2017-01-02.

[Merceur 2016] F. Merceur, T. Carval, J.J.H. Buck, T. Loubrieu and S. Pouliquen: DOIs for ocean data. General principles and selected examples. Ifremer report IMN/IDM/ISI/FM/16-024, May 13 2016. https://dx.doi.org/10.13155/44515.

[Moats 1997] R. Moats: RFC 2141 URN Syntax. Available at https://www.rfc-editor.org/rfc/rfc2141.txt. Accessed 2017-01-30.

[Myers 2015] Myers, J; M. Hedstrom; D. Akmon; S. Payette; B. A. Plale; I. Kouper ; S. McCaulay; R. McDonald; I. Suriarachchi; A. Varadharaju; P. Kumar; M. Elag; J. Lee; R. Kooper and L. Marini, Towards sustainable curation and preservation, in Proc. IEEE eScience Conf. 2015, 526-535. https://doi.org/10.1109/eScience.2015.56

[OASPA 2017a] Open Access Scholarly Publishers Association. 2017. OASPA | Open Access Scholarly Publishers Association. http://oaspa.org/ Accessed 2017-01-04.

[OASPA 2017b] Open Access Scholarly Publishers Association. 2017. Frequently Asked Questions – OASPA (FAQ1: Does membership of OASPA require a specific type of license?). http://oaspa.org/information-resources/frequently-asked-questions/#FAQ1 Accessed 2017-01-04.

[OASPA 2017c] Open Access Scholarly Publishers Association. 2017. Members – OASPA. http://oaspa.org/membership/members/ Accessed 2017-01-04.

[OASPA 2017d] Open Access Scholarly Publishers Association. 2017. Mission and Purpose – OASPA. http://oaspa.org/about/mission-and-purpose/ Accessed 2017-01-04.

[OpenAIRE 2017a] OpenAIRE. 2016. Project Factsheet | General Information. https://www.OpenAIRE.eu/project-factsheets Accessed 2017-01-05.

[OpenAIRE 2017b] OpenAIRE. 2016. OpenAIRE 2020 Work Packages, Tasks, Deliverables. https://www.OpenAIRE.eu/OpenAIRE2020-wps-tasksAccessed 2017-01-05.

[OpenAIRE 2017c] OpenAIRE. 2017. OpenAIRE — Search publications, datasets, projects… https://www.OpenAIRE.eu/search/find?keyword= Accessed 2017-01-05.

[OpenAIRE 2017d] OpenAIRE. 2017. OpenAIRE — Guidelines for Data Archives – OpenAIRE Guidelines 3.0 documentation. https://guidelines.openaire.eu/en/latest/data/index.html Accessed 2017-01-05.

[OpenAIRE 2017e] OpenAIRE. 2017. 1. Identifier (M) – OpenAIRE Guidelines 3.0 documentation. https://guidelines.openaire.eu/en/latest/data/field_identifier.html Accessed 2017-01-05.

[OpenAIRE 2017f] OpenAIRE. 2017. OpenAIRE | FAQ. Question 27: “What can I do to ensure and improve OpenAIRE compatibility?”. https://www.openaire.eu/support/faq#ifaqCat-19 Accessed 2017-01-05.

[ORCID 2016a] ORCID. 2016. Our vision. http://orcid.org/node/8 Accessed 2016-12-19.

[ORCID 2016b] ORCID. 2016. Welcome to Collect & Connect: ORCID’s integration and engagement program. http://orcid.org/content/collect-connect Accessed 2016-12-19.

[ORCID 2016c] ORCID. 2016. ORCID member organizations. http://members.orcid.org/member-list Accessed 2016-12-19.

[ORCID 2016d] ORCID. 2016. ORCID statistics. https://orcid.org/statistics Accessed 2016-12-19.

[ORCID 2016e] ORCID. 2016. Requiring ORCID in Publication Workflows: Open Letter. http://orcid.org/content/requiring-orcid-publication-workflows-open-letter Accessed 2016-12-19.

[ORCID 2016f] ORCID. 2016. Frequently asked questions: What is the relationship between ISNI and ORCID? https://orcid.org/faq-page#n80 Accessed 2016-12-19.

[Parsons 2010] M.A. Parsons, R.E. Duerr and J.-B. Minster, ”Data citation and peer review”, EOS, *Transactions of the American Geophysical Union* vol 91, no 34, 24 August 2010, 297-304. http://dx.doi.org/10.1029/2010EO340001

[Pentz 2016] Pentz, Ed: The Organization Identifier Project: a way forward. Blog published 2015-10-31. http://blog.crossref.org/2016/10/the-oi-project.html Accessed 2016-12-19.

[PR Newswire 2016] PR Newswire. Acquisition of the Thomson Reuters Intellectual Property and Science Business by Onex and Baring Asia Completed. http://www.prnewswire.com/news-releases/acquisition-of-the-thomson-reuters-intellectual-property-and-science-business-by-onex-and-baring-asia-completed-300337402.html Accessed 2016-12-19.

[Rauber 2015] A. Rauber et al., “Data citation of evolving data. Recommendations of the Working Group on Data Citation (WGDC)”. Preliminary report from 20 Oct 2015. Available at https://rd-alliance.org/system/files/documents/RDA-DC-Recommendations_151020.pdf Accessed 2017-01-30.

[Rauber 2016] A. Rauber, A. Asmi, D. van Uytvanck and S. Pröll, ”Identification of Reproducible Subsets for Data Citation, Sharing and Re-Use”. *Bulletin of IEEE Technical Committee on Digital Libraries*, vol. 12, issue 1, May 2016, 6-15. Available at http://students.cs.tamu.edu/ldmm/tcdl/v12n1/papers/IEEE-TCDL-DC-2016_paper_1.pdf

[Redhead 2015] Redhead, C. 2015. DET – interest levels rising – OASPA. Blog post published 2015-06-24. http://oaspa.org/det-interest-levels-rising/ Accessed 2017-01-04.

[ResearchGate 2016a] ResearchGate. 2016. Generating a DOI. https://explore.researchgate.net/display/support/Generating+a+DOI Accessed 2016-12-20.

[ResearchGate 2016b] ResearchGate. 2016. Adding research. https://explore.researchgate.net/display/support/Adding+research Accesed 2016-12-20.

[RDA 2016a] Research Data Alliance. 2016. Who is RDA?. https://www.rd-alliance.org/about-rda/who-rda.html Accessed 2016-12-19.

[RDA 2016b] Research Data Alliance. 2016. Interest groups. https://www.rd-alliance.org/groups/interest-groups Accessed 2016-12-19.

[RDA 2016c] Research Data Alliance. 2016. Working groups. https://www.rd-alliance.org/groups/working-groups Accessed 2016-12-19.

[RDA Europe 2016] Research Data Alliance Europe. 2016. RDA Europe. https://europe.rd-alliance.org/ Accessed 2016-12-19.

[Schindler 2008] Schindler, U, & Diepenbroek, M. 2008. Generic XML-based framework for metadata portals, *Computers & Geosciences*, 34(12)12, December 2008, pp.1947-1955, http://dx.doi.org/10.1016/j.cageo.2008.02.023.

[Scholix 2016] Scholix. 2016. RDA and ICSU-WDS Announce the Scholix Framework for Linking Data and Literature. Press release published 2016-06-20. https://www.icsu-wds.org/news/press-releases/scholix-framework-for-linking-data-and-literature/at_download/file1/Press_Release_SCHOLIX_June2016.pdf Accessed 2017-01-30.

[Schwardmann 2015] U. Schwardmann, "ePIC Persistent Identifiers for eResearch" Presentation at the joint DataCite-ePIC workshop Persistent Identifiers: Enabling Services for Data Intensive Research, Paris, 21 Sept 2015. Available at https://doi.org/10.5281/zenodo.31785.

[Socha 2013] Y.M. Socha, ed., "Out of cite, out of mind: The current state of practice, policy, and technology for the citation of data". *Data Science Journal* vol. 12, 13 Sept 2013. Available at https://www.jstage.jst.go.jp/article/dsj/12/0/12_OSOM13-043/_pdf Accessed 2017-01-30.

[Starr 2015] J. Starr, E. Castro, M. Crosas, M. Dumontier, R.R. Downs, R. Duerr, L.L. Haak, M. Haendel, I. Herman, S. Hodson, J. Hourclé, John Ernest Kratz, J. Lin, L. Holm Nielsen, A. Nurnberger, S. Pröll, A. Rauber, S. Sacchi, A. Smith, M. Taylor and T. Clark: Achieving human and machine accessibility of cited data in scholarly publications. *PeerJ Comput. Sci.* 1:e1; DOI http://dx.doi.org/10.7717/peerj-cs.1

[Stehouwer 2014] H. Stehouwer and P. Wittenburg, eds. "Second year report on RDA Europe Analysis Programme: Survey of EU Data Architectures", Deliverable D2.5 from the RDA Europe project (FP7-INFRASTRUCTURES-2012-1), 2015. Available at https://rd-alliance.org/sites/default/files/Survey%20of%20data%20mangement%20needs.docx Accessed 2017-01-30.

[Stocker 2016] M. Stocker: Persistent Identification of Instruments. Presentation at the PIDapalooza meeting in Reykjavik, Iceland, November 9-10, 2016. https://doi.org/10.6084/m9.figshare.4246100.v1.

[Stockhause 2015] M. Stockhause, F. Toussaint and M. Lautenschlager: "CMIP6 Data Citation and Long-Term Archival", CMIP6 project report. Available via link on the WIP Resources page https://www.earthsystemcog.org/projects/wip/resources/ Accessed 2017-01-30.

[STM Publishers 2017a] International Association of Scientific, Technical and Medical Publishers. 2017. About the Association. http://www.stm-assoc.org/about-stm/about-the-association/ Accessed 2017-01-04.

[STM Publishers 2017b] International Association of Scientific, Technical and Medical Publishers. 2017. Now available: SCHOLIX. http://www.stm-assoc.org/standards-technology/free-stm-webinars-learn-scholix/ Accessed 2017-01-04.

[Uhlir 2012] P.F. Uhlir, rapporteur, "For Attribution — Developing Data Attribution and Citation Practices and Standards". Summary of an international workshop (August 2011), National Research Council, 2012. Available at http://www.nap.edu/openbook.php?record_id=13564. Accessed 2017-01-30.

[Web of Science 2016a] Web of Science. 2016. Data Citation Index — IP & Science — Thomson Reuters. http://wokinfo.com/products_tools/multidisciplinary/dci/ Accessed 2016-12-19.

[Web of Science 2016b] Web of Science. 2016. The repository selection process — IP & Science — Thomson Reuters. http://wokinfo.com//products_tools/multidisciplinary/dci/selection_essay/?utm_source=false&utm_medium=false&utm_campaign=false Accessed 2016-12-19.

[Weigel 2014] T. Weigel, T. DiLauro and T. Zastrow, "RDA PID Information Types Working Group: Final Report", Final report from the Research Data Alliance PID Information Types (PIT) Working Group, released on 2014-11-25, 25pp, http://dx.doi.org/10.15497/FDAA09D5-5ED0-403D-B97A-2675E1EBE786.

[Weigel 2016] T. Weigel, M. Lautenschlager, P. Wittenburg, L. Lannom and D. van Uytvanck: RDA Data Fabric Configurations — Thoughts about PID Centric DMA: Towards a global Virtualisation Layer. Available at https://www.rd-alliance.org/sites/default/files/PIDCDMA-extension-v4.docx. Accessed 2017-01-10.

[Wikipedia 2017a] Wikipedia: Content negotiation. https://en.wikipedia.org/wiki/Content_negotiation. Accessed 2017-01-30.

[Wikipedia 2017b] Wikipedia: Google Scholar. https://en.wikipedia.org/wiki/Google_Scholar Accessed 2017-01-04.

[Wikipedia 2017c] Wikipedia: OpenURL. https://en.wikipedia.org/wiki/OpenURL Accessed 2017-01-05.

[Wikipedia 2017d] Wikipedia: SFX (software). https://en.wikipedia.org/wiki/SFX_(software) Accessed 2017-01-05.

[Wikipedia 2017e] Wikipedia: PANGAEA. https://en.wikipedia.org/wiki/PANGAEA_(data_library) Accessed 2016-12-15.

[Wilkinson 2016] M. D. Wilkinson, M. Dumontier, I.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J-W. Boiten, L. B. da Silva Santos, P.E. Bourne, J. Bouwman, A.J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C.T. Evelo, R. Finkers, A. Gonzalez-Beltran, A.J.G. Gray, P. Groth, C. Goble, J.S. Grethe, J. Heringa, P.A.C. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S.J. Lusher, M.E. Martone, A. Mons, A.L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M.A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao and B. Mons: The FAIR Guiding Principles for scientific data management and stewardship, *Sci. Data* 3:160018 doi: http://dx.doi.org/10.1038/sdata.2016.18.

[Zenodo 2016a] Zenodo. 2016. About Zenodo. https://zenodo.org/about Accessed 2016-12-20.

[Zenodo 2016b] Zenodo. 2016. FAQ. https://zenodo.org/faq Accessed 2016-12-20.

[Zenodo 2016c] Zenodo. 2016. Features. https://zenodo.org/features Accessed 2016-12-20.

[Zhao 2015] Z. Zhao: The theme of data for science. Presentation at the 1st ENVRIPLUS week meeting, Prague, The Czech Republic, November 2015.