D6.2
A report on negotiations with publishers, providers of existing data citation systems and other scientific organisations on implementing a global data citation system

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

LEADING BENEFICIARY: LUND UNIVERSITY

Author(s):	**Beneficiary/Institution**
Margareta Hellström (lead), Maria Johnsson	LU (Lund University, ICOS)
Frank Toussaint, Stephan Kindermann	DKRZ (IS-ENES)
Dan Lear	MBA (EMBRC)
Robert Huber, Markus Stocker	UniHB (University of Bremen)
Ingemar Häggström, Carl-Fredrik Enell	EISCAT (EISCAT Scientific Association)

Accepted by: Alex Vermeulen (WP6 leader) and Zhiming Zhao (Theme 2 leader)

Deliverable type: [REPORT]

Dissemination level: PUBLIC

Deliverable due date: 30.04.2018/M36

Actual Date of Submission: 30.04.2108/M36

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 654182

ABSTRACT

This deliverable reports the group efforts of Work Package 6 Task T6.1 during M24 to M36 to prepare and initiate a dialogue ("negotiations") with publishers, providers of existing data citation systems and other scientific organisations on raising awareness of what environmental and climate research infrastructures view as essential identification and citation-related services that are required in order to reach the ultimate goal of a "global data citation system". The activities include creating a network of contacts with a number of actors across the identification and citation landscape, organising a workshop bringing these actors together with ENVRI partners in order to exchange information on current practices, and undertaking a survey aimed at examining the attitudes of relevant publishers and PID, citation & indexing service providers towards citation-related issues identified as important by ENVRI partners.

REPORT REVIEWERS

Project internal reviewer(s):

Project internal reviewer(s):	**Beneficiary/Institution**
Leonardo Candela (Theme 2 expert)	CNR
Markus Stocker (Theme 2 member representative)	UniHB

DOCUMENT VERSION HISTORY

Date	**Version**
1.11.2017	Outline available for comments
1.3.2018	First draft of outline
4.4.2018	Version sent for internal review
26.4.2018	Version sent to Theme 2 and WP 6 leaders
30.4.2018	Final version, submitted by project office

DOCUMENT AMENDMENT PROCEDURE

Amendments, comments and suggestions should be sent to the lead author (Margareta Hellström margareta.hellstrom@nateko.lu.se)

TERMINOLOGY

Acronyms and specialist terminology used in this report are explained in **Appendix A**.
In addition, a complete ENVRIplus project glossary is provided online here:
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 2015a]

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 harmonize 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.

EXECUTIVE SUMMARY

This second deliverable of ENVRIplus Work Package 6 (WP6), "Inter RI data identification and citation services", is concerned with 1) identifying challenges standing in the way of Environmental Research Infrastructures (ENVRIs) as they move towards implementing comprehensive citation and referencing of entities (data, instruments, samples, etc.) related to their activities; 2) defining relevant "negotiation" partners among publishers, persistent identifier service providers, citation indexers and other organisations; 3) initiating a constructive and positive dialogue with these actors; and 4) and feeding back the outcomes and results of the discussions into both the ENVRIs' own practices as well as those of their end users, and the global research data science community.

In support of especially points 2 and 3, WP6 has created a network of contacts with a number of organisations, and initiated a discussion with these on a range of topics. A workshop was organised in October 2017, bringing together representatives from research infrastructures (RIs) with technical specialists from e.g. publishers, data repositories and service providers. High-priority issues were identified -- including citation of non-data objects, referencing of dynamic data, methods to pinpoint subsets of larger datasets, and management of data collections.

Building on the workshop outcomes, a questionnaire-driven survey was performed, aimed at mapping out the views and stance on the high-priority issues. Interestingly, the survey responses indicated that while there was consensus on basic issues like supporting the use of PIDs also for non-data entities and the need to provide long-term sustainable services, the views on how to best manage citations of data subsets, data collections and dynamic data were much more fragmented, with no clear trends between the various respondent categories.

A useful network of contacts has now been established, and the "negotiation" activities met with great interest from the publishers, PID service providers and indexers who participated in the workshop and the survey. The concrete outcomes – including a clearer understanding of which citation and identification-related issues are of high priority to the ENVRI community – will now feed back into the further work of WP6 towards designing and implementing services addressing those issues still remaining in the way of achieving comprehensive and trustworthy identification and citation practices for Earth Science researchers in Europe and globally.

TABLE OF CONTENTS

1 ABOUT WORK PACKAGE 6

The overarching objective of ENVRIplus Work Package 6 (WP6), "Inter RI data identification and citation services", is to improve the efficiency of data identification and citation in the environmental and Earth science fields by providing access to convenient, effective and interoperable identifier management and citation services. This WP highlights identification and citation in environmental RIs, reviews available technologies and develops common services for these operations. In addition, it aims to set up a dialogue between ENVRI community partners and relevant actors and organisations involved in the provisioning of services related to identification and subsequent citation of digital representations of objects from all stages of the research life cycle.

The first WP6 deliverable (D6.1, [Hellström 2017]), summarised the associated technological needs and requirements of the ENVRIplus partners, outlined a suggested common system design for Identification and Citation, and mapped the landscape of publishers, PID service providers and other actors in the scholarly data management and curation world.

This second WP6 deliverable -- A report on negotiations with publishers, providers of existing data citation systems and other scientific organisations on implementing a global data citation system -- continues where D6.1 left off, addressing the following points from the WP6 Description of Work [ENVRIplus 2015b]:

- Perform an analysis of the latest statuses of 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.
- Promote the needs of environmental RIs in the global context. Once the RIs have decided their priorities, these should be addressed to initiatives targeting pan-European Digital Identifier e-infrastructures as well as global initiatives such as the Belmont Forum and the Research Data Alliance. The goal of respective agreements should be a widely accepted and supported model.
- 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.

Finally, in parallel to the above activities WP6 is also concerned with the development of a number of use case studies. The outcomes of two of these, a) the implementation of an on-line, standards-based publication mechanism for marine biological data and b) the development of a workflow and guidance for citation tracking models, will be reported in the final deliverable D6.3. Two additional use cases are managed together with WP9, and have been described and reported in deliverable D9.1 [Chen 2017].

2 UPDATING THE LANDSCAPE: EXTERNAL INITIATIVES

Since the publication of ENVRIplus deliverable D6.1 [Hellström 2017], there have been a lot of activities around various aspects of (data) citation. Some projects, like THOR, have now come to an end, while others including FREYA are just starting. Here we summarise some of the most important recent initiatives that are likely to influence the data citation practices of ENVRIplus members and their end user communities.

2.1 FORCE11

The international community for scholarly communication FORCE11[1] is engaged in several questions concerning research data and citation [FORCE11 2018]. The community, established in 2011, has several working groups in this area:

- Data Citation Implementation Pilot (DCIP)
- Resource Identification Technical Specifications Working Group
- Software Citation Implementation Working Group

FORCE11 engages publishers, libraries, researchers, research funders, and many other groups with an interest in scholarly communication, and the community is behind many initiatives in research data and citation. The work within FORCE11 is of high relevance to ENVRIplus and should be monitored regularly. FORCE11 may also be a good discussion partner for the continuing work in ENVRIplus WP6.

2.2 European Open Science Cloud (EOSC)

Formed as result of a summit in June 2017, the European Open Science Cloud[2] (EOSC) [EOSC 2018] involves many different organisations such as research infrastructures, funders, research institutions, and commercial providers in developing systems and infrastructures for Open Science. The EOSC Declaration [EOSC 2017] describes in detail what needs to be done in different aspects of Open Science, including a section on the need for setting up a data citation system to reward the provision of open data. This statement is sure to catalyse the continuing work in ENVRIplus and its RIs. The projects EOSCpilot[3] and EOSC-hub[4] will contribute to implement the visions in EOSC.

2.3 Scholix

The Scholix[5] project [Burton 2017] is progressing with publishers, data repositories and other organisations signing up to the Scholix link exchange service. For ENVRIplus, it is essential to follow the developments in Scholix, as they are working with interoperability and standards for linking literature with research data. Scholix could therefore be a good discussion partner for ENVRIplus WP6.

2.4 Make Data Count[6]

Measuring data usage will be critical for future global citation practices [Kratz 2015]. The more researchers claim credit for their research data, the more there will be a need to establish metric systems for data, which is the focus of the project Make Data Count [Make Data Count 2018a]. The early first phase (2014-2015) of the project was described in the previous ENVRIplus WP6 deliverable. The project addresses is now addressing the social and technological barriers of widespread incorporation of data-level metrics in research data management systems.

With partners from research, libraries, funders, and publishers the project addresses is working on the following four main objectives:

---

[1] https://www.force11.org/
[2] http://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud
[3] https://eoscpilot.eu/
[4] https://www.eosc-hub.eu/
[5] http://www.scholix.org/
[6] https://makedatacount.org/

1. to publish a new COUNTER code of practice regarding how data usage should be measured and reported;
2. to deploy a data-level metrics (DLM) aggregation and publication service based on the open-source Lagotto software and hosted by DataCite;
3. to integrate the DLM service with new data sources and clients;
4. to perform advocacy and training regarding the importance and use of DLM [Make Data Count 2018b]

2.5 Metadata 2020

Common standards for metadata and interoperable systems by which metadata are transferred correctly will by all means have an impact on the future practice of, and systems for, global data citation. Metadata 2020[7] is a major initiative engaging in these issues. It is a collaborative network and initiative working on the matters of richer, connected and reusable metadata for all research outputs. The network aims to create awareness and resources for all stakeholders involved in metadata. Currently it runs projects concerning research communication, metadata recommendations, defining terms about metadata, etc. The network involves researchers, publishers, librarians, repositories, service providers, and funders.

2.6 Citation Style Language

The Citation Style Language (CSL)[8] is an open XML-based language to describe the formatting of citations and bibliographies. Reference management programmes using CSL include Zotero[9], Mendeley[10] and Papers[11]. It was created for integration with the OpenOffice.org application suite and can be used to transfer different citation styles. A style repository containing more than 9000 styles of which more than 1700 are unique is supported by the CSL project.

2.7 Project THOR

The THOR project[12] was identified in the previous WP6 deliverable (see chapter 7.2.1 in [Hellström 2017]) as an important contributor to forming European practices for applying PIDs in the research life cycle. The project, which ended in November 2017, aimed at ensuring that every researcher, at any phase of their career, will have seamless access to PIDs for their research artefacts.

As both ENVRIplus and THOR have identified PID services as central pieces of technology, there were many possible topics for exchange of ideas and contacts, starting with a collaboration on developing the concept of persistent identification of instruments. The idea was presented at the first PIDapalooza conference in 2016[13], where we underscored the importance of metadata about instruments to science. The topic was also presented again later at the Helsinki meeting, this time to an audience of environmental RI specialists. Given the interest by several of the attendees, we discussed the idea of forming a Research Data Alliance working group (WG), and over the next months the "RDA WG Persistent Identification of Instruments"[14] was established. The new WG had its kick-off meeting at the RDA 11th Plenary Meeting in March 2018.

---

[7] http://www.metadata2020.org/
[8] http://citationstyles.org/ and http://citationstyles.org/authors/
[9] https://www.zotero.org/
[10] https://www.mendeley.com/
[11] https://www.readcube.com/papers/
[12] https://project-thor.eu/
[13] https://pidapalooza.org/index2016.html
[14] https://www.rd-alliance.org/groups/persistent-identification-instruments

In addition, THOR and ENVRIplus co-organised a workshop[15] in Helsinki (March 2017), looking at different ways for environmental RIs to integrate ORCID[16] IDs and services in their data management structures. Out of 17 ENVRI partner RIs, 12 indicated after the workshop that they planned to do so. As an example, the ICOS Carbon Portal Authentication Service has now integrated ORCID identity-based login. We suggest that more can be done by ENVRI partners to e.g. integrate ORCID beyond login, in particular to facilitate setting up cross-links between contributors and published data or other kinds of relevant entities, and then to share such cross-link information with the PID infrastructure (e.g. by adding relevant DOI record metadata in the case of establishing ORCID-DataCite cross-links).

2.8 Project FREYA

Following on the experiences of the previous EU-funded projects THOR and ODIN[17], the new project FREYA[18] will continue and work on an extended infrastructure for PIDs, to improve discovery, navigation, retrieval, and access to research resources. The FREYA project engages people from PID service providers, publishers, research infrastructures etc., and it works closely with both RDA and EOSC [FREYA 2018a]. As a vision, FREYA has established the three following pillars:

- The PID Graph connects and integrates PID systems, creating relationships across a network of PIDs and serving as a basis for new services
- The PID Forum promotes engagement with the global community via the RDA and through organising conferences, workshops and other PID-themed events
- The PID Commons addresses the sustainability of the PID infrastructure resulting from FREYA beyond the lifetime of the project [FREYA 2018b].

Representatives from several ENVRIplus partners have recently expressed interest in joining the FREYA Ambassador program, aimed at helping to spread information on best practices for PID use across end user communities, and it is hoped that the exchange of information and experiences will be extended even further.

3 TOWARDS A GLOBAL CITATION SYSTEM

In this chapter we introduce the concept of "global citation system", and provide a brief background of the current thinking about identification and citation. We then conclude by outlining a set of challenges related to citation and publishing of data faced by ENVRI community members.

3.1 Introduction

As stated in the WP6 Description of Work [ENVRIplus 2015b], environmental RIs are often built on a large number of distributed observational or experimental sites, run by hundreds of scientists and technicians, financially supported and administered by a large number of institutions. It is clearly expected by the stakeholders that the data produced by the ENVRI community is distributed in a FAIR manner and under open access policies – but at the same time, it is also crucial that all downstream use of the data can be appropriately quantified and that proper credit is given to the institutions and individuals involved in the production, quality assurance and dissemination activities.

[15] https://project-thor.eu/2017/04/13/envrid-integrating-orcid-ids-in-environmental-research-infrastructures/
[16] http://orcid.org
[17] https://cordis.europa.eu/project/rcn/105189_en.html
[18] https://www.project-freya.eu/en

Achieving these goals will require giving data producers access to a set of easy-to-use services for

- data storage in accredited repositories with associated metadata catalogues,
- registration of persistent identifiers,
- citing and referral to individual data objects or collections of these in a standardised manner, and
- collection of citation and usage statistics, down to the level of individual data objects or even subsets of these
- linking between data, metadata, people, places, instrumentation, samples, scientific articles & other literature

The combination of such services, suitable for application at all stages of the research data lifecycle (from raw data all the way to derived data products), can be thought of as a "citation system". Because of the variability in standards and practices between scientific domains and disciplines [Martone 2014], and even between countries, regions and continents, it is questionable whether it makes sense to strive towards a single, all-encompassing and world-wide citation system that would fit all users and usage scenarios. However, if we focus on the disciplines falling under the Earth Science heading, it may be possible to find flexible-enough solutions that can bridge across different requirements and traditions, and achieve a "global" citation system for the research domains represented by the ENVRI Community: atmosphere, biosphere, marine and solid earth.

3.2 Background

In this section we provide brief summaries of some important aspects of identification and citation. More in-depth information can be found in other ENVRIplus outputs, including deliverables D5.1 [Atkinson 2016] (especially chapters 2.3.1, 3.2 and 4.1.2) and D6.1 [Hellström 2017] (chapters 2, 3 and 4).

3.2.1 Identification

Unequivocal identification of resources and objects underlies all aspects of today's research data management. The assignment of persistent and unique identifiers (PIDs) to digital objects and other resources, and to simultaneously store specific metadata (url, originator, type, date, size, checksum, etc.) in the PID registry database, is a first and indispensable step towards ensuring reproducibility of research [Duerr 2011], [Stehouwer 2014], [Almas 2015].

3.2.2 Citation

A second, equally necessary step is the subsequent use of consistent and standardised methods to use the PIDs to cite or refer to digital objects (or digital representations of physical entities) wherever it is required [Atkinson 2016]. Indeed, a pervasive adoption of persistent identifiers in research is expected to contribute significantly to quality and efficient re-use of research data, by increasing the overall efficiency of the research process and by enhancing the interoperability between RIs, ICT service providers and users [Almas 2015]. In addition, 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].

3.2.3 Hesitancy to share data

However, a number of surveys have indicated that many data producers (both individual scientists and research groups of varying sizes) are hesitant to share their data openly, mainly due to a perceived lack of proper attribution of data [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.

Easy and ubiquitous access to services that support identification and citation throughout the entire research life cycle are therefore vital towards persuading data creators of the value of sharing their data and metadata, and convincing data users of the need to cite data and other research entities in a way that allows assignment of scientific credit to the producers which – in a wider context - – will help convince funders to continue to support data gathering and curation.

3.2.4 Dividing the responsibilities

Following on from the technology review reported in deliverable D5.1 ([Atkinson 2016]) and the subsequent work reported in deliverable D6.1 ([Hellström 2017]), there are a number of basic research data life cycle "components" that are required to support proper citation of data.

As repositories mostly rule the vocabulary for metadata description, they need to ensure that all elements of the metadata schemata they use are clearly described. To ensure that there are no misunderstandings, especially when PID registry metadata are shared with other services or transferred to other communities, it is especially important that metadata concepts relating to roles (author, contributor, owner etc.) are unambiguously described and that guidance is provided for both the type of content and acceptable formats.

Data providers must make sure that all metadata relevant to supporting the assignment of persistent identifiers are made available to the repository and/or the PID service, preferably already at the time of ingestion or registration, including a complete list of all individuals (and institutions) that should be able to receive credit for the work invested into collecting, quality assessing and finalizing of the data object(s).

Repositories must ensure that their cataloguing systems are ready to capture all relevant metadata, to disseminate these e.g. via landing pages, and to format them in a way that enables correct harvesting by other portals or metadata stores. They must also provide the possibility to assign persistent identifiers (PIDs) of a suitable type (depending on e.g. the research community involved) to the data sets, and in doing so transfer all relevant metadata (kernel) information into the PID registry's database. Repositories should also make available suitable citation strings or similar for each data set.

End users of data sets must take care to store all metadata that are associated with research data they download in such a way that there is no ambiguity of the original sources, including creator(s). When writing up and publishing outcomes of their research, e.g. as articles or reports, they must ensure that every data set used is properly cited, for example by using citation strings suggested by the repository, provided these are correctly formatted. In addition, any *derived* data set associated with such a scientific publication should have relevant provenance included in its own metadata, preferably generated (quasi-)automatically during the data analysis process.

3.2.5 Dynamic data, subsets and collections

Environmental observational data 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]. Finding standards for citing subsets of potentially very large

and complex data sets poses a special problem, as outlined by Huber *et al*. [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 data set 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.

Furthermore, there is a growing trend to create collections of research-related items that have some common theme or characteristic. Examples include lists of individual data objects that belong together, packages of data and associated metadata, and more complex "research objects" [Wf4Ever 2013] that may also contain annotations, related articles and reports, etc. Collections can be defined by the original producers, but may also be collated by end users of the data – and may thus contain information from a large variety of sources and types. 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 [Weigel 2017].

3.2.6 Actionable links to data

To make the links "actionable" (especially for automated workflows executed by machines without human intervention), it must be possible to either arrive directly at the cited data object itself, or at least to unambiguously extract the data object link from the information available at the URL that the PID resolves to [Socha 2013]. This has several aspects: if the PID resolves to a "landing page", this should support content negotiation and/or follow recommended standards for how to include machine-actionable links [Starr 2015]. There is a lot of confusion on how to "URL-ify" a DOI or PID in order to make it machine-actionable, and the nuances and differences between pre-pending e.g. http://dx.doi.org, http://doi.org, or https://doi.org to a (Handle System) PID are often not recognized [Wittenburg 2017]. Even worse, there are numerous examples of citations to data containing links pointing directly to the URL of the data object or its landing page – not recognizing or realizing that if this URL becomes invalid (also known as "link rot"), the data will no longer be retrievable.

3.3 ENVRIplus challenges on the way to proper citation and data publishing practices

Many RIs may experience problems or difficulties in their role as data producers or data publishers, in particular when metadata describing citation-relevant information about data sets are exchanged between different portals or cataloguing systems. In ENVRIplus WP6, we have encountered a number of cases that illustrate these problems. In the following, we describe some in more detail.

3.3.1 Issues with defining roles for authors, contributors and editors

As the main person that should get credit for publishing a scientific data set is the data creator, special care is to be taken to ensure a proper assignment.

The classical concept of "author" refers to the person (or persons) who has or should have intellectual property rights regarding a specific work, such as an article or in this context, a data set. Unlike this, for a data collection it is instead typically a coordinator, editor or data collector that should receive production-related merits rather than content related, although he/she

mostly has the responsibility on the data package as a whole. This is why this person usually is marked by "(Ed.)" behind the name, for example see the reference recommendation on page ii of the 2014 IPCC report [IPCC 2014].

A problem arises, when as in the database world - maybe triggered by schemas like Dublin Core - the person responsible for a collection is often referred to as the database's "author", despite a lack of scientific intellectual contributions. An example in point is the Springer Verlag report "Earth System Modelling – Volume 6" [Budich 2013], where only two of the three individuals acknowledged as "authors" actually contributed to the book's content. Thus, the category "creator" does not allow distinguishing between author and editor as long as the creator is not given a “role”. In a further example, taken from the climate data domain, it proved impossible to assign a DOI to a large data collection consisting of a number of individual datasets with in total hundreds of contributors (real "authors"), as the collection coordinator refused to have his name listed in the "author" metadata field of the collection object's PID registry entry. This illustrates that the general concept of “creator” urgently needs a role for a more specific declaration of possible merits.

3.3.2 Formatting citations based on PID registry metadata

Another example of problems that may occur in data transfer between portals is that the portals have different policies/practice for their citation formats. “Portal 1”, which is the data publisher, has a fairly detailed citation format in which information on data contributors is included. However, when the same data are transferred and made accessible on another data portal, “Portal 2”, the (default) citation format may be completely different and the contributor information is no longer visible or even available. As a result, when users display the data on “Portal 2” they may not get the full information about the original contributors. The extreme number of different citation formats led to the development of CrossCite’s DOI Citation Formatter[19], a tool which was elaborated by a consortium of DataCite, Crossref, and mEDRA together with a Chinese PID registration agent[20]. Here one can select from far more than 1500 different publishers’ formats. This service, which also features a machine-actionable API, could prove useful for both repositories and individual ENVRI community members wishing to provide citation strings for their data products.

3.3.3 Citation and provenance information get lost between publishing portals

Transferring data from one publishing portal to another may cause problems with capturing essential information on provenance when data are delivered to other publishing portals. As an example, we discuss the case in which “organisation A” has offered access to its data to the discipline-specific portal of “organisation B”. The two organisations agreed to deliver the data according to a specific protocol and in a specific format. “Organisation B” offers access to the data via its homepage. However, on this website the data were not published according to the citation style of “organisation A”, which lead to the unfortunate situation that original author information was lost. Furthermore, as the same data were transferred from “organisation B” to a third publishing portal, here “organisation C”, the provenance information of “organisation A” was lost and the only way to recognize the original publisher was via DOI. For the authors of the data, this could lead to ‘misassigned’ or even 'lost' credit for data they produced. Furthermore,

[19] https://citation.crosscite.org/
[20] http://www.doi.org.cn/

those who re-use the data do not have correct information about the authors. This example demonstrates the importance of setting standards for metadata to be accompanied with data when transferred from one publishing portal to another.

3.3.4 Metadata for experiments such as configuration, scheduling and measurement modes

Some ENVRIplus RIs run experiments that are very flexible with respect to scheduling of operation and configuration of measurement modes, being able to run several experiments in parallel. Here, metadata and provenance information describing the scheduling and configuration of the experiments play an important role. While these metadata can be expressed according to the ENVRI Reference Model [ENVRI RM V2.1 2016], other domain specific implementations such as XML data files, SQL databases, and HDF5 file properties should also be part of the configuration metadata. Integrating information on software and code in the metadata descriptions of the experiments would also enhance the repeatability of the experiments [Smith 2016].

3.3.5 GBIF and the integrated publishing toolkit

The development of the use case for the standardised description of marine biodiversity data was enabled by established tools developed by the Global Biodiversity Information Facility (GBIF). The Integrated Publishing Toolkit is a Java based free, open source application for the publication and sharing of biodiversity data. The tool natively supports the ability to automatically generate and assign a DOI to the described data through integration with DataCite[21] or EZID[22]. Working with the publishers of the Marine Biodiversity Records journal[23] it was intended to provide a mechanism for the capture of the underlying geospatial distributional data alongside the peer reviewed publication. The journal deals with changes in the geographical range of marine species, including the effects and impacts of invasive species and responses to climate change. The data underpinning these papers are therefore of high societal and research value and their capture and integration into national, regional and global infrastructures is important.

Marine Biodiversity Records is published by Springer, who classifies journals and their associated data policies in four progressively stringent categories: 1) Data sharing and data citation is encouraged but not required; 2) Data sharing and evidence of data sharing are both encouraged; 3) Data sharing is encouraged and statements of data availability are required; and 4) Data sharing, evidence of data sharing and peer review of data are all required. The Marine Biodiversity Records journal is currently classified as Type 3, the full policy for which states *"The journal strongly encourages that all datasets on which the conclusions of the paper rely should be available to readers. We encourage authors to ensure that their datasets are either deposited in publicly available repositories (where available and appropriate) or presented in the main manuscript or additional supporting files whenever possible"*

However, despite strong encouragement in the policies, detailed guidance in the instructions to authors and the provision of a standards-based template, the uptake of the GBIF data management services has been extremely poor. Unless mandated by funders or strict journal policies, there is arguably still insufficient motivation for authors to consider the archiving of the underlying data.

[21] http://datacite.org
[22] https://ezid.cdlib.org/
[23] https://mbr.biomedcentral.com/

On a more positive note, the marine domain benefits from well-established PIDs for the description of taxonomy and geographic and administrative areas. The World Register of Marine Species (WoRMS[24]) aims to provide a comprehensive and authoritative list of names for marine species, including synonyms. WoRMS has chosen the AphiaID identifier system, which is widely used across the marine domain. Whilst less ubiquitous, the MarineRegions.org and associated MRID identifier serve to provide a standard, relational list of geographic names, coupled with information and maps of the geographic location of these features.

3.3.6 The lack of a "d index"

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 sizable 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 — we may call it 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[25] is a non-profit organisation that operates the services that allow 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". It will require an joint effort by data providers, DataCite and other PID infrastructure, 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 essential avant-garde role in showcasing such a development.

3.4 Moving forward and finding solutions

The awareness of the importance of incorporating persistent identifiers into their operational research data management is growing amongst ENVRIplus partners, but conversations during collaboration meetings with RI representatives not engaged in WP 6 indicate that most project partners are implementing their own, customized solutions – often covering only parts of the "best practices" outlined in deliverable D6.1 [Hellström 2017]. In addition, as evidenced from responses to the survey conducted in preparation of ENVRIplus deliverable D5.1 [Atkinson 2016] (see also [ENVRI Community 2016]), many ENVRI partners appear to have a quite vague

[24] http://www.marinespecies.org/about.php
[25] http://www.datacite.org

knowledge of how often their end user communities are citing data, and what standards (if any) are being used.

To address these problems, ENVRIs will certainly need to adjust their own data management practices – but in order to do so, they are to a large extent dependent on the availability of suitable identification and citation-related services. However, as evidenced by the list of challenges outlined above in Chapter 3.3, there are significant gaps between the functionalities currently on offer and the specific needs and requirements of the ENVRI Community members, especially regarding how to:

- Consistently define roles for authors, contributors and editors in the metadata of DOI records
- Use this information to create appropriate citation strings
- Ensure proper exchange of attribution metadata between portals and repositories
- Collect usage statistics and other bibliometric information also for data
- Define methods and standards for managing dynamic data, subsets and collections

4 NEGOTIATIONS WITH PUBLISHERS AND OTHER ACTORS

We note that although the term negotiation is defined as e.g. "a process in which two or more parties resolve a dispute or come to a mutual agreement" [Merriam-Webster 2018], many associate the term with commercial interests, as in settling the monetary value, or requiring the *de facto* exchange, of goods or services. We have therefore preferred to consistently use the words 'dialogue' and 'discussion' in our contacts with representatives from the various organisations that were invited to participate in the activities reported in this deliverable.

4.1 Motivation

While it is very difficult for individual scientists or even RIs to contact publishers, PID registries, citation indexers and other service providers in order to raise awareness of these important issues, the ENVRI community stands a much larger chance to initiate a meaningful dialogue and to ultimately bring about the desired changes – and it is with this conviction that WP6 has been given the mission to engage with publishers and other actors reported on in this chapter. By negotiating with these parties, we aim to:

- set up and maintain a network of contacts to support this dialogue
- make their requirements and wishes of ENVRI Community members known to all the relevant actors
- learn about the current developments being undertaken by publishers, service providers and indexers
- organise workshops, surveys and other activities during which discussions and collaboration around common projects can take place
- find a common agreement on what functionalities need to be improved or added
- establish a priority plan for the required developments
- clearly define how to share the responsibilities between developers (service providers) and testers (ENVRI community)

4.2 Identifying the discussion partners

The first task of the "negotiation" process has therefore been to identify the relevant discussion partners, which should ideally include representatives from all of the following categories:

- Organisations that bring together professionals interested in research data management must be made aware of the needs and requirements of environmental RIs
- 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
- Data aggregators and data portal publishers should set up mechanisms suitable to preserve those original data citations provided by data archives and avoid overwriting these with own citation patterns and identifiers.
- Agencies that collect references and citations need to provide comprehensive bibliometrics, altmetrics and statistics that allow to distinguish data usage from traditional article citations

Based on our earlier review of the service provider and publisher "landscape" (see Chapter 6 of D6.1 [Hellström 2017]), a list of potential organisations and companies was compiled. A special effort was made to identify relevant contact persons within each organisation, ideally with good knowledge not only of technical details but also of the broader, strategic and/or policy-related picture. This proved to be quite a challenge, as a majority of the organisations on our list did not provide detailed contact information on their web sites, or only offered general communication channels such as Twitter or Facebook.

4.3 Setting a timeline for the "negotiations"

We note that this deliverable should be regarded as a progress report, rather than a final summary of all negotiation-related WP6 activities. Indeed, this is not the end point of the dialogue as such, and we aim to continue to exchange views and information with the negotiation partners up to the end of the ENVRIplus project and beyond.

The following timeline[26] indicates the order and duration of the negotiation-related WP6 activities:

- Workshop with publishers, PID providers, indexers and ENVRIplus RIs — October 2017
- Shortlist of negotiation partners, and finalised strategy ready — December, 2017
- Questionnaire to negotiation partners — January-February 2018
- Summarizing results in Deliverable D6.2 (this document) — April 2018
- Follow-up of questionnaire outcomes with negotiation partners — May-July 2018
- Revisit the RI data identification & citation requirements survey — August-October 2018
- Report back to ENVRI community — November 2018

4.4 Negotiation activities

In this section, we summarise the activities we have undertaken so far in support of the dialogue between ENVRIplus partners and our discussion partners. The main achievements include holding a workshop and carrying out a questionnaire-based survey. In addition, WP6 members have taken active part in relevant working groups and discussion fora organised by e.g. the RDA. Follow up negotiations between data archives (PANGAEA) and data aggregators and data portals such as GBIF and EUDAT (B2FIND service) are ongoing in order to ensure the RIs interests as described above and to preserve original data citation patterns and identifiers.

[26] Note that this has been somewhat revised with respect to what was previously stated in D6.1 ([Hellström 2017]).

4.4.1 Workshop

On October 18, 2017 WP6 organised the workshop "Closing the gap: The need for tools to identify, track and cite environmental research data". The event, hosted by the German Climate Computing Centre (DKRZ) in Hamburg, Germany, was designed to bring together publishers, PID service providers and environmental research communities to discuss common challenges.

Participants and program

The workshop was attended by 22 participants from the ENVRIplus partner RIs, publishers, PID service providers and experts on PID systems. The workshop started off with some presentations of use cases from ENVRIplus RIs, which were followed by presentations on PIDs for non-data research objects and by presentations from PID service providers and data publishers. The workshop was finished by common discussion. **Appendix B** contains the program of the workshop, while **Appendix C** lists all presentations including abstracts.

Summary of the topical discussions

The last point of the workshop was a common discussion on the following topics:

Dynamic data, including versioning

This topic is already being discussed in different ENVRIplus communities. Arguably, ENVRIplus should focus more on the next generation of recommendations and reference implementations. Are we, ENVRI community, happy with these technologies, specifically DOI, DataCite, ePIC? The notion of "campaign data" was also discussed. How to define a "campaign"?

Sites, instruments, samples

This topic focused on the citation of non-data objects, such as instruments, sites and samples. The discussion was on how to create links to software and instruments. The notion of linking to a "site" was also discussed, as more and more ENVRIplus RIs are introducing this concept. What should constitute a "site" within the ENVRIplus communities?

Management of data collections

We discussed how to give credit to those who contribute with data to others' data collections. Could this be solved in a good and practical manner? Is there a WG, e.g. within RDA, investigating this problem/matter? We also discussed the process of creating a data collection with data from different sources. What "role" should that creator take in that case, e.g. the editor role?

Sustainable data typing services

Under this topic we looked at the activities of RDA and ISO regarding data typing services. There does not appear to exist any current use cases from within the European environmental and climate research domains for how data typing could be consistently applied.

4.4.2 Questionnaire/survey

In order to get a grasp of what different organisations may offer in terms of services and support on PIDs, ENVRIplus WP6 performed a survey directed to publishers, PID service providers, data usage indexers and other organisations with an interest on data citation and PIDs. The questionnaire consisted of 10 questions and was distributed in January 2018 to 39 organisations. We received 18 responses in total, i.e. 7 answers from publishers, 6 answers from PID service providers, 2 answers from data usage indexers, and 3 answers from other organisations. See

Appendix D for a list of the organisations who answered the questionnaire. A majority of the responses were detailed and elaborated, giving us confidence that the material would allow us to obtain an overview of the current views of the participating organisations.

Questions & answers

Here follows a summary of the results of the questionnaire arranged by the 10 questions asked. Note that each of the four major target groups (*Publishers, PID service providers, Indexers* and *Others*) received slightly different versions of the survey; they differed in the order and included questions. See **Appendix E** for excerpts from selected responses (slightly modified to protect respondent anonymity).

Q1: The concept "persistent" in persistent identifiers, what does that mean to you and the services in your organization?

Across the board of respondents, many answered that PIDs will resolve to stable and unique entities, so there is a clear consistency here. Some also brought up the need for maintenance of technological infrastructure and consistent rules for identifiers. There was no particular difference across the target groups.

Q2: What types of PIDs may you allow in your services?

The participants were presented with the following options (with brief explanations of each):

- ARKs (Archival Resource Keys)
- DOIs (Digital Object Identifiers)
- Handles (Handle System)
- LSID (Life Science Unique Identifiers)
- PURL (Persistent URL)
- URN (Uniform Resource Name)
- Others

There was no clear trend on the usage of PID types in the target groups; some supported several PID types, while others only allowed one or two. Across the responses, the distribution of the usage of PID types was as follows: DOIs (17), Handles (11), PURLs (8), URNs (8), ARKs (4), LSIDs (4) and other PID types (7). It should be noted that DOIs – which in this context were specified as PIDs provided by DataCite - are a sub-category of the more generic Handle System, which is used by many other PID service providers such as ePIC.

Q3: What is your opinion on harvesting records with PIDs from scientific sources? Do you have solutions or technologies for this in your organization?

This question was included to the target-group *Indexing organisations* only, and one of their feedbacks was that it is highly useful to harvest records with PIDs from scientific sources. It facilitates the quality check of the source and the curation of that record.

Q4: What is your opinion on PID based references pointing to samples, instruments and stations in scientific articles, e.g. PIDs to non-data objects? Would it be feasible to support PID services to these references to non-data objects?

A clear majority (15 respondents) were positive to the idea of PIDs pointing to samples, instruments and stations. Everybody was also positive to support PID services to these references to non-data objects. The answers from *Publishers* and *Other organisations* were more on a general concept level, while the answers from *PID service providers* and *Indexing*

organisations were more technical. Some of the PID service providers also gave examples of PIDs pointing to samples, instruments and stations.

Q5: What is your opinion on peer-review of datasets to obtain a PID?

Judging by the answers given, this question was a bit unclear to the target groups. There was confusion whether it was about peer-review in order to obtain a PID for a dataset or whether it regarded peer-review of records containing PIDs as a kind of quality check. The opinions differed among the survey participants. Some *Publishers* and *PID service providers* were positive to peer-review of datasets while others were quite negative and did not see the meaning or value of it.

Q6: What is your opinion on allowing bibliographic references to be made to dataset fragments or subsets (i.e. by appending pointer information to the PID of the dataset)? Do your services support pointers to subsets in bibliographic references?

10 of the respondents were positive to the idea of granularity in referencing to data sets but were uncertain how this would work in practice with rules and standards. Most of *Publishers* were positive to the idea of references pointing to fragments or subsets. *PID service providers*, were also positive and gave more descriptive and technical answers. There might be other ways of demonstrating which data set fragments have been used than "building" it in a PID solution, such as showing it through the metadata.

Q7: What is your opinion about using PIDs for data collections (i.e. collections of several datasets)? Do your services support PIDs for data collections?

The opinions were diverse. Several declared that this is a good practice and should be supported while others viewed difficulties, i.e. ensuring credits to individual contributions in large data collections. One participant stated that data collections might be helpful for journal publishers since they could reduce the need to incorporate large numbers of citations of individual datasets.

Q8a: Do your services support bibliographic references for dynamic data sets?

This question was only directed to *Publishers*, *PID service providers* and *Indexing organisations*, and their feedback was quite mixed. A total of 8 out of 13 answered that they have some kind of service or support for bibliographic references to dynamic data sets. Other organisations were presented with the alternative question "What is your opinion of dynamic data sets?" and had several comments such as the importance of adding necessary context and metadata in conjunction with the dynamic data sets.

Q8b: What is your opinion on assigning PIDs to search queries rather than assigning PIDs to the results of a query?

The responses we received to this question differed quite a lot. Only a few were positive to the idea of assigning PIDs to search queries, while others saw risks with such a practice, e.g. the query returning different results or returning multiple results that would surely cause confusion. One comment was that this would require a rigorous versioning of databases to ensure the same search results were returned. Other comments indicated that PIDs should point to durable or specific objects, while search queries are somewhat evanescent rather than persistent.

Q9a: In relation to persistence of PIDs, what is your opinion on the "sustainability" of your products and services?

All target groups declared their products and services to be sustainable, in particular those products connected to DOIs. Membership structure, integration in scholarly publishing

workflows, solid business models are important factors for the sustainability of the products and services.

Q9b: What time frame would constitute “sustainable” for your services?

Most of the respondents had long perspective of more than 10 years of sustainability of their products and services, often even longer: 75, 100 years or indefinitely.

Q10: Do you have any other comments or ideas on persistent identifiers and data citation?

Some of the *Publishers* brought up the importance of establishing standards for PIDs and mentioned initiatives such as Scholix as essential. Some from the other target groups mentioned factors such as the management of PIDs and metadata enrichment as important. According to one comment, there is also a challenge in communicating to journals that authors need to be allowed to cite data sets in the same way as they cite papers.

4.4.3 Other contact points

In addition to the workshop and the survey, WP6 representatives are also engaging in discussions and dialogue on identification and citation-related issues in a number of other forums. It may be noted that such information exchange forms an integral part of the data management development work of the RIs, and is to a large extent taking place in parallel to the ENVRIplus-funded activities. This is especially true for the engagement in RDA groups.

Related RDA groups & discussions

In the RDA's own words [Research Data Alliance 2018], it "provides a neutral space where its members can come together through focused global Working and Interest Groups to develop and adopt infrastructure that promotes data-sharing and data-driven research, and accelerate the growth of a cohesive data community that integrates contributors across domain, research, national, geographical and generational boundaries."

RDA currently has 61 Interest Groups (IGs), which members are experts from the community that are committed to directly or indirectly enabling data sharing, exchange, or interoperability, and 32 Working Groups (WGs), comprised of experts from the international community engaged in creating deliverables that will directly enable data sharing, exchange, or interoperability.

Over ten individuals working for ENVRIplus partners, including of course also institutions associated with WP6, are members of RDA and actively engaged in IGs and WGs that are concerned with topics and issues with a strong coupling to identification and citation:

- Persistent Identification of Instruments WG[27]
- PID Kernel Information WG[28]
- Research Data Collections WG[29]
- Data Versioning WG[30]
- Data Fabric IG[31]
- Persistent Identifiers IG[32]

[27] https://www.rd-alliance.org/groups/persistent-identification-instruments
[28] https://www.rd-alliance.org/groups/pid-kernel-information-wg
[29] https://www.rd-alliance.org/groups/research-data-collections-wg.html
[30] https://www.rd-alliance.org/groups/data-versioning-wg
[31] https://www.rd-alliance.org/group/data-fabric-ig.html
[32] https://www.rd-alliance.org/groups/pid-interest-group.html

Conferences and topic meetings

There have been several meetings and conferences on PIDs and data citation, where representatives from ENVRIplus Work Package 6 have been present and followed the discussions.

Events organised by Research Data Alliance

In addition to plenary meeting sessions organized by the working and interest groups mentioned above, RDA has organised several events dedicated to persistent identifiers and data citation the last years. Two particular events are worth mentioning here, namely:

- Persistent Identifiers: Enabling Services for Data Intensive Research, organised as a pre-RDA workshop by DataCite and ePIC on 21 September 2015 in Paris[33].
- Views about PID Systems, RDA Europe training and workshop held in Munich on 31 August – 2 September 2016.[34]

PIDapalooza meeting series

The "PIDapalooza" meetings are solely dedicated to technologies and services around persistent identifiers. It was first held in 2016 in Reykjavik, Iceland, with a second event organised in Girona, Spain in January 2018 [35]. The latter event attracted many participants from service providers, libraries, publishers and research institutes, who came together to discuss and debate PIDs from a multitude of perspectives, such as how to organise the workflow for assigning PIDs, what metadata schemes to use at PID registry level, and new emerging uses of PIDs for people, sensor platforms and instruments, and physical samples.

4.4.4 Follow-up activities

The publication of this deliverable does not mark the conclusion of the WP6 "negotiation" activities. Several follow-up activities are being planned for the final year of ENVRIplus.

Follow-ups with survey participants

All respondents to the survey questionnaire indicated that they wanted to receive a copy of the deliverable, once finished. A majority expressed a willingness to continue the dialogue on citation and identification issues, also beyond the scope of ENVRIplus.

Revisiting ENVRIplus partner requirements

In the framework WP5, a survey of all ENVRIplus partner RIs was performed in late 2015-early 2016, with the aim to map out their technical requirements related to research data management (see deliverable D5.1 [Atkinson 2016]). We suggest to get back in contact with all respondents in order to find out if their understanding and insights into data identification & citation have significantly changed since then and/or whether their needs for services and support are now different. (The 2015-2016 responses are available via links in [ENVRI Community 2016]).

---

[33] https://www.rd-alliance.org/group/pid-ig/post/persistent-identifiers-enabling-services-data-intensive-research.html and https://blog.datacite.org/recap/

[34] https://www.rd-alliance.org/views-about-pid-systems-training-course-and-workshop-31-august-2-september-2016-garchingmunich, https://www.rd-alliance.org/sites/default/files/attachment/20160901_RDA_PID_event_Garching_report_final.pdf

[35] https://pidapalooza.org/

Identifying expertise on identification & citation in the wider ENVRI Community

In addition, it will be crucial to the continued WP6 work to identify and maintain an up to date list of all Identification & Citation experts in the infrastructures, projects and other organisations that make up the wider ENVRI Community. Access to such a list will facilitate information exchange, as well as streamline efforts to perform future surveys and training activities, also beyond the end of the ENVRIplus project.

5 OUTCOMES AND CONCLUSIONS

This second deliverable of ENVRIplus Work Package 6 (WP6), "Inter RI data identification and citation services", is concerned with 1) identifying challenges standing in the way of Environmental Research Infrastructures (ENVRIs) as they move towards implementing comprehensive citation and referencing of entities (data, instruments, samples, etc.) related to their activities; 2) defining relevant "negotiation" partners among publishers, persistent identifier service providers, citation indexers and other organisations; 3) initiating a constructive and positive dialogue with these actors; and 4) and feeding back the outcomes and results of the discussions into both the ENVRIs' own practices as well as those of their end users, and the global research data science community.

In support of especially points 2 and 3, WP6 has created a network of contacts with a number of organisations, and initiated a discussion with these on a range of topics. A workshop was organised in October 2017, bringing together representatives from research infrastructures (RIs) with technical specialists from e.g. publishers, data repositories and service providers. High-priority issues were identified -- including citation of non-data objects, referencing of dynamic data, methods to pinpoint subsets of larger datasets, and management of data collections.

Building on the workshop outcomes, a questionnaire-driven survey was performed, aimed at mapping out the views and stance on the high-priority issues. Interestingly, the survey responses indicated that while there was consensus on basic issues like supporting the use of PIDs also for non-data entities and the need to provide long-term sustainable services, the views on how to best manage citations of data subsets, data collections, and dynamic data were much more fragmented, with no clear trends between the various respondent categories.

A useful network of contacts has now been established, and the "negotiation" activities met with great interest from the publishers, PID service providers and indexers who participated in the workshop and the survey. The concrete outcomes – including a clearer understanding of which citation and identification-related issues are of high priority to the ENVRI community – will now feed back into the further work of WP6 towards designing and implementing services addressing those issues still remaining in the way of achieving comprehensive and trustworthy identification and citation practices for Earth Science researchers in Europe and globally.

6 IMPACT ON PROJECT

The outcomes of the "negotiations" reported here will need to be taken into account during the finalisation of the development of services performed as part of WP6 in particular and Theme 2 in general:

- Attribution of the data contributors is of vital importance for RIs. Integrating an accurate data citation system in the data workflow is therefore critical.

- Increased awareness among the ENVRI partners of the current thoughts and views on identification & citation issues that are held by important actors, including publishers, PID service providers and citation indexers
- Insights that will inform the way that ENVRIplus partners and the ENVRI community use persistent identifiers
- Feedback into cataloguing and metadata-related activities, especially concerning information to be used for defining citation strings

7 IMPACT ON STAKEHOLDERS

It is our hope that this report will provide stakeholders – including the leaderships of the involved RIs, their respective funding agencies as well as relevant national agencies and policy makers – with:

- a snapshot of (data) citation related questions and issues currently seen as important by ENVRIs
- the corresponding views and opinions of the publishers, PID service providers and other actors
- an outlook of future activities intended to support the dialogue and discussions towards a global data citation system

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the contributions of all the individuals and organisations who participated in the workshop and responded to our questionnaire. We also would like to thank professor Michael Witt of Purdue University, USA for reading an early draft of the deliverable and sharing his views on data citation with us.

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.

[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

[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. https://doi.org/10.5281/zenodo.34542.

[Budich 2013] R. Budich, R. Redler. Earth System Modelling – Volume 6, ESM Data Archives in the Times of the Grid. Springer. 2013. https://doi.org/10.1007/978-3-642-37244-5.

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

[Chen 2017] Y. Chen, B. Grenier, M. Hellström, A. Vermeulen, M. Stocker, R. Huber, B. Magagna, I. Häggström, M. Fiebig, P. Martin, D. Vitale, G. Judeau, T. Carval, T. Loubrieu, A. Nieva, K. Jeffery, L. Candela and J. Heikkinen: Service deployment in computing and internal e-Infrastructures. ENVRIplus Deliverable 9.1, submitted on August 31, 2017. Available at http://www.envriplus.eu/wp-content/uploads/2015/08/D9.1-Service-deployment-in-computing-and-internal-e-Infrastructures.pdf.

[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.

[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 Community 2016] Identification and citation requirements. Article in the ENVRI Collaboration and Documentation wiki, https://wiki.envri.eu/display/EC/Identification+and+citation+requirements. Page version of May 25, 2016. Accessed on April 24, 2018.

[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.

[ENVRIplus 2015a] ENVRIplus project description, public part. ENVRIplus Grant Agreement, Annex 1, part B. Horizon 2020 project no. 654182. Associated with document Ref. Ares(2015)1488547. Available at http://www.envriplus.eu/wp-content/uploads/2015/08/ENVRIplus_PartB_public.pdf.

[ENVRIplus 2015b] ENVRIplus Description of Work (DoW), public part. ENVRIplus Grant Agreement, Annex 1, part A. Horizon 2020 project no. 654182. Associated with document Ref. Ares(2015)1488547. Available at http://www.envriplus.eu/wp-content/uploads/2015/08/ENVRIplus_DoW_public.pdf.

[EOSC 2017] EOSC Declaration, European Open Science Cloud, New Research & Innovation Opportunities, Brussels 26 October 2017. Available at https://ec.europa.eu/research/openscience/pdf/eosc_declaration.pdf. Accessed on March 29, 2018.

[EOSC 2018] European Open Science Cloud. Available at https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud. Accessed on March 29, 2018.

[FORCE 11 2018] FORCE 11. Available at https://www.force11.org/. Accessed on March 29, 2018.

[FREYA 2018a] FREYA. Available at https://www.project-freya.eu/en. Accessed on March 29, 2018.

[FREYA 2018b] FREYA, About. Available at https://www.project-freya.eu/en/about/mission. Accessed on March 29, 2018.

[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. http://dx.doi.org/10.1007/s12145-014-0202-2.

[Hellström 2017] M. Hellström, M. Lassi, A. Vermeulen, R. Huber, M. Stocker, F. Toussaint, M. Atkinson and M. Fiebig: A system design for data identifier and citation services for environmental RIs projects to prepare an ENVRIPLUS strategy to negotiate with external organisations. ENVRIplus Deliverable D6.1, submitted on January 31, 2017. Available at http://www.envriplus.eu/wp-content/uploads/2015/08/D6.1-A-system-design-for-data-identifier-and-citation-services-for-environmental-RIs.pdf

[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. http://dx.doi.org/10.6084/m9.figshare.1285728

[IPCC 2014] IPCC. Climate Change 2013 The Physical Science Basis, Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Available at http://www.climatechange2013.org/images/report/WG1AR5_Frontmatter_FINAL.pdf . Accessed on April 25, 2018.

[Kratz 2015] J. Kratz and C. Strasser. Making data count. Sci. Data 2:150039 (2015). http://dx.doi.org/10.1038/sdata.2015.39.

[Make Data Count 2018a] Make Data Count. Available at https://makedatacount.org/. Accessed on March 29, 2018.

[Make Data Count 2018b] About, Make Data Count. Available at https://makedatacount.org/about/ . Accessed on April 24, 2018.

[Martone 2014] M. Martone ed., Joint Declaration of Data Citation Principles, Data Citation Synthesis Group and FORCE11, San Diego CA, 2014. Available at https://www.force11.org/group/jointdeclaration-data-citation-principles-final.

[Merriam-Webster 2018] Negotiation. A Merriam-Webster Online Dictionary article, available at https://www.merriam-webster.com/dictionary/negotiation. Accessed on March 28, 2018.

[Rauber 2015] A. Rauber, A. Asmi, D. van Uytvanck, S. Pröll. Data Citation of Evolving Data, Recommendations of the Working Group on Data Citation (WGDC). Available at https://www.rd-alliance.org/groups/data-citation-wg.html . Accessed on 25 April 2018.

[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

[Research Data Alliance 2018] Who is RDA? https://www.rd-alliance.org/about-rda/who-rda.html. Accessed on March 29, 2018.

[Smith 2016] A.M. Smith, D.S. Katz, K.E. Niemeyer and FORCE11 Software Citation Working Group. Software citation principles. PeerJ Computer Science, vol 2:e86, 2016. http://dx.doi.org/10.7717/peerj-cs.86.

[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.

[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; 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://rdalliance.org/sites/default/files/Survey%20of%20data%20mangement%20needs.docx.

[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.

[Weigel 2017]. T. Weigel, B. Almas, F. Baumgardt, Th. Zastrow, U. Schwardmann, M. Hellström, J. Quinteros and D. Fleischer. Recommendation on Research Data Collections. Draft report from the RDA research data collection working group (October 19, 2017). Available via https://github.com/RDACollectionsWG/specification/blob/master/Recommendation%20package/rda-collections-recommendation.pdf.

[Wf4Ever 2013] Workflow4Ever project. Wf4Ever Research Object Model 1.0, Wf4Ever Specification 30 November 2013. Available at http://wf4ever.github.io/ro/.

[Wittenburg 2017] P. Wittenburg, M. Hellström and C.-M. Zwölf (Eds.). Persistent identifiers: Consolidated assertions. Status of November, 2017. Zenodo. http://doi.org/10.5281/zenodo.1116189

APPENDIX A. ACRONYMS AND SPECIAL TERMS

This appendix is based on the official ENVRI terminology and glossary, as available at the ENVRI community wiki site (see https://wiki.envri.eu/pages/viewpage.action?pageId=14452608).

A.1. Terminology & glossary specific to this deliverable

ARK: Archival Resource Keys, a type of persistent identifier.

CrossRef: Non-profit membership organization making research outputs easy to find, cite, link, and assess.

CSL: Citation Style Language, an open XML-based language used to format citations and bibliographies.

DataCite: Global non-profit organisation that provides persistent identifiers (DOIs) for research data.

Data collection: A number of datasets grouped together as one entity.

DLM: Data-Level Metrics, an aggregation and publication service developed by Scholix

DO: Digital Object.

DOI: Digital Object Identifier.

Dynamic data: Refers to datasets that may change over time, e.g. because new data has been added, updates or changes of data have been made.

ePIC: European Persistent Identifier Consortium.

EZID: Service from the California Digital Library allowing to create and manage long-term globally unique IDs for data and other sources.

FORCE11: international community for scholarly communication.

Fragment dataset: A specific subset of a larger dataset.

FREYA: European research project, funded by Horizon 2020. Follow-up of the THOR project.

Handles: Short for the Handle System, a type of persistent identifier.

HDF5: Hierarchical Data Format (HDF) is a set of file formats (HDF4, HDF5) designed to store and organize large amounts of data.

LSID: Life Science Unique Identifiers, a type of persistent identifier.

mEDRA: Multilingual European Registration Agency for DOI persistent identifiers for any form of intellectual property on a digital network.

Metadata 2000: Initiative to define common standards for metadata interoperability.

ORCID: Non-profit organization providing unique identifiers for researchers.

PID: Persistent digital identifier.

Pidapalooza: Series of international events dedicated to technologies and services around persistent identifiers.

PURL: Persistent URL, a type of persistent identifier.

Query store: Instead of storing many duplicates of subsets of data it is possible to create specific queries in order to identify and obtain certain subsets of data. The queries may be stored in a query store, enabling re-use.

Scholix: Scholarly link exchange is a high level interoperability framework for exchanging information about the links between scholarly literature and data, as well as between datasets.

SQL: Structured Query Language, a domain-specific language used in programming and designed for managing data held in a relational database management system.

THOR: European research project, funded by Horizon 2020. Precursor to FREYA.

URL: Uniform Resource Locator, a location-based uniform resource identifier.

URN: Uniform Resource Name, a type of persistent identifier.

WoRMS: The World Registry of Marine Species.

XML: Extensible Markup Language.

A.2. Other technical terms and acronyms used in ENVRIplus deliverables

API: Application Program Interface, is a set of routines, protocols, and tools for building software applications

Biodiversity: is the variety of different types of life found on earth

Biodiversity metrics: measurements of the number of species and how they are distributed

CERIF: Common European Research Information Format

CIARD RING: A global directory of information services and datasets in agriculture

D4Science: is an organisation offering a Hybrid Data Infrastructure service and a number of Virtual Research Environments

Data stream: is a sequence of digitally encoded coherent signals used to transmit or receive information that is in the process of being transmitted

Data pipeline: In computing, a pipeline is a set of data processing elements connected in series, where the output of one element is the input of the next one.

DCAT: is a resource description format vocabulary designed to facilitate interoperability between data catalogues

DIRAC: Distributed Infrastructure with Remote Agent Control. High-Throughput computing platform operated by EGI.

EduGAIN: is an international inter-federation service interconnecting research and education identity federations

E-infrastructure: can be defined as networked tools, data and resources that support a community of researchers, broadly including all those who participate in and benefit from research

FIM4R: Federated Identity Management for Research collaborations

gCube: is an open-source software toolkit used for building and operating Hybrid Data Infrastructures enabling the dynamic deployment of Virtual Research Environments by favouring the realisation of reuse oriented policies

HPC: High Performance Computing

HTC: High Throughput Computing

IoT: The Internet of Things - is a scenario in which objects, animals or people are provided with unique identifiers and the ability to transfer data over a network without requiring human-to-human or human-to-computer interaction.

ICT: Information and Communications technology

IG: Interest Group, open-ended topic group, for example in the Research Data Alliance

IPR: Intellectual Property Rights

KOS: Knowledge Organization Systems - is a generic term used in Knowledge organization about authority lists, classification systems, thesauri, topic maps, ontologies etc.

LOD: Linked open data is linked data that is open content

LOV: Linked Open Vocabularies

Metadata: is data that describes other data. Metadata summarizes basic information about data, which can make finding and working with particular instances of data easier

NGI: National Grid Initiative

NMI: National Metrological Institutes

NREN: National Research and Education Network

NRT: Near Real Time - refers to the time delay introduced, by automated data processing or network transmission, between the occurrence of an event and the use of the processed data (For example, a near-real-time display depicts an event or situation as it existed at the current time minus the processing time, as nearly the time of the live event)

ODP: 1) Open Distributed Processing (for the ENVRI Reference Model); 2) Online Data Processing

OIL-E: The Open Information Linking model for Environmental science - is a semantic linking framework

Ontology: (In computer science and information science) an ontology is a formal naming and definition of the types, properties, and interrelationships of the entities that really or fundamentally exist for a particular domain of discourse

QoE: Quality of user experience

Over dispersion: a statistical characteristic of data such that the data have more clusters than compared to what might be expected if the data were distributed randomly in proportion to the time/space available.

NetCDF: a file format.

OceanSITES: a worldwide system of long-term, open-ocean reference stations measuring dozens of variables and monitoring the full depth of the ocean from air-sea interactions down to the seafloor

OOI: Ocean Observatories Initiative

RDA: Resource Description and Access, a standard for descriptive cataloguing. See also A.3. (Organisational acronyms) below.

RM: Reference Model - is an abstract framework or domain-specific ontology consisting of an interlinked set of clearly defined concepts produced by an expert or body of experts in order to encourage clear communication

SensorML: The primary focus of the Sensor Model Language is to provide a robust and semantically-tied means of defining processes and processing components associated with the measurement and post-measurement transformation of observations

Semantics: is the study of meaning

Syntax: In computer science, the syntax of a computer language is the set of rules that defines the combinations of symbols that are considered to be a correctly structured document or fragment in that language

SLA: Service Level Agreement

SME: Small and medium-sized enterprise

UV: Unmanned vehicles

VL: Virtual Laboratory

VRE: Virtual Research Environments, web based package tailored to a specific community

WG: Working Group, time-limited topic group, for example in the Research Data Alliance

A.3. Organisational acronyms

ACTRIS: Aerosols, Clouds, and Trace gases Research Infrastructure network. ENVRIplus partner.

AnaEE: Analysis and Experimentation on Ecosystems. European research infrastructure, ENVRIplus partner.

AQUACOSM: EU network of mesocosms facilities for research on marine and freshwater ecosystems open for global collaboration

CDI: Collaborative Data Infrastructure. European e-service provider organisation,

CEA: Commissariat à l'Energie Atomique et aux Energies Alternatives. French research agency, ENVRIplus participant.

CINECA: Consorzio Interuniversitario. Italian non-profit research consortium, ENVRIplus participant.

CNR: Consiglio Nazionale delle Richerche. Italian national research council, ENVRIplus participant.

CNRS: Centre National de la Recherche Scientifique. French research organisation, ENVRIplus participant.

CODATA: Committee on Data for Science and Technology.

ConnectinGEO: Coordinating an Observation Network of Networks EnCompassing saTellite and IN-situ to fill the Gaps in European Observations

COOPEUS: Strengthening the cooperation between the US and the EU in the field of environmental research infrastructures. Project funded under EU FP7, continued as COOP+ under Horizon 2020.

COPERNICUS: previously known as GMES (Global Monitoring for Environment and Security), is the European Programme for the establishment of a European capacity for Earth Observation

CREEM: Centre for Research into Ecological and Environmental Modelling, operated by University of St Andrews (USTAN).

CSC: Center for Science (Tieteen tietotekniikan keskus Oy). Finnish national high-performance computing centre, ENVRIplus participant.

CU: Cardiff University. UK university, ENVRIplus participant.

DANUBIUS: The international center for Adavanced studies on river-sea systems

DASSH: Data archive for seabed species (a UK marine biology resource centre)

DiSSCo: Distributed Systems of Scientific Collections

DKRZ: German Climate Computation Center (Deutsches Klimarechenzentrum GmBH). German research organisation, ENVRIplus participant.

EAA: Umweltbundesamt GmbH - Environment Agency Austria. Austrian governmental agency, ENVRIplus participant.

EEA: European Environment Agency

EGI: Stichting European Grid Initiative. European research foundation, ENVRIplus participant.

EISCAT: EISCAT Scientific Association. European research organisation, ENVRIplus participant.

EISCAT_3D: Multi-static phased array radar system. Operated by EISCAT Scientific Association, ENVRIplus partner.

EMBL: European Molecular Biology Laboratory. European research organisation, ENVRIplus participant.

EMBRC: European Marine Biological Resource Centre. A research infrastructure and consortium of research organisations interested in marine biology. ENVRIplus partner.

EMODNET: The European Marine Observation and Data Network

EMRP: European Metrology Research Programme

EMSC: European-Mediterranean Seismological Centre. European non-governmental organisation, ENVRIplus participant.

EMSO: European Multidisciplinary Seafloor and Water Column Observatory. European research infrastructure, ENVRIplus partner.

EOSC: European Open Science Cloud. Initiative from the European Commission.

EPOS: The European Plate Observing System. European research infrastructure, ENVRIplus partner.

ERIS: Environmental Research Infrastructure Strategy 2030

ESONET VI: European Seafloor Observatory NETwork. European research infrastructure, ENVRIplus partner.

ETHZ: Eidgenössische Technische Hochschule Zürich. Swiss technical university, ENVRIplus participant.

EUDAT: H2020 project on Research Data Services, Expertise & Technology Solutions (previously funded by FP7). Continues as the Collaborative Data Infrastructure (CDI).

EUFAR: European Facility for Airborne Research

EURO-ARGO: European research infrastructure, ENVRIplus partner.

EUROCHAMP2020: European atmospheric simulation chambers

EUROFLEETS: New operational steps towards an alliance of European research fleets. ENVRIplus partner.

EUROGOOS: European Global Ocean Survey System. International non-profit association, ENVRIplus participant.

EuroSITES: European Ocean Observatory Network

FixO3: Fix point open ocean observatories (survey programme). European research infrastructure, ENVRIplus partner.

FMI: Finnish Meteorological Institute (Ilmatieteen Laitos). Finnish research and service agency, ENVRIplus participant.

FZJ: Research Centre Jülich (Forschungszentrum Jülich GmbH). German research centre, ENVRIplus participant.

GBIF: Global Biodiversity Information Facility

GEO: The Group on Earth Observations.

GEOMAR: Helmholtz Zentrum für Ozeanforschung Kiel. German research institution, ENVRIplus participant.

GEOSS: Global Earth Observation System of Systems, coordinated by GEO (The Group on Earth Observations)

GMES: Global Monitoring for Environment and Security, previous name for COPERNICUS.

GROOM: Gliders for research ocean observation and management

HELIX Nebula: partnership between big science and big business in Europe that is charting the course towards the sustainable provision of cloud computing - the Science Cloud

IAGOS: In-service Aircraft for a Global Observing System. European research infrastructure, ENVRIplus partner.

ICOS: Integrated Carbon Observation System. European research infrastructure, ENVRIplus partner.

ICSU: The International Council for Science

IFREMER: Institute Français de Recherche Pour l'Exploitation de la Mer. French research organisation, ENVRIplus participant.

INGV: Istituto Nazionale di Geofisica e Vulcanologia. Italian research institute, ENVRIplus participant.

INRA: Institut National de la Recherche Agronomique. French research institute, ENVRIplus participant.

INSPIRE: Integrated Sustainable Pan-European Infrastructure for Researchers in Europe

INTERACT: International Network for Terrestrial Research and Monitoring in the Arctic. European research infrastructure, ENVRIplus partner.

IPBES: Intergovernmental Platform on Biodiversity & Ecosystem Services

IS-ENES: Infrastructure for the European Network for Earth System Modelling. European research infrastructure, ENVRIplus partner.

JERICO: Towards a joint European research infrastructure network for coastal observatories. European research project, ENVRIplus partner.

LifeWatch: European e-Science infrastructure for biodiversity and ecosystem research. ENVRIplus partner.

LTER: The Long-term Ecological Research Network. International research organisation.

LTER-Europe: European Long-term Ecosystem Research network of 21 national LTER networks. ENVRIplus partner.

LU: Lund University (Lunds universitet). Swedish university, ENVRIplus participant.

MARUM: Centre for Marine Environmental Sciences at University of Bremen (UniHB).

MBA: Marine Biological Association of the United Kingdom. UK research organisation, ENVRIplus participant.

NERC: Natural Environment Research Council. UK research council, ENVRIplus participant.

NILU: Norwegian Institute of Air Research (Norsk Institutt for Luftforskning). Norwegian research institute, ENVRIplus participant.

OASIS: Advancing Open Standards for the Information Society (non-profit consortium)

PANGAEA: Information system and data publisher for geoscientific and environmental data, operated by MARUM and UniHB. German data repository, ENVRIplus participant.

PLOCAN: Oceanic Platform of the Canary Islands (Consorcio Para el Diseno, Construccion, Equipamiento y Explotacion de la Plataforma Oceanica de Canarias). Spanish research organisation, ENVRIplus participant.

RCN: Research Council of Norway (Norges Forskningsrad). Norwegian national research council, ENVRIplus participant.

RDA: Research Data Alliance. International organisation working to promote collaboration on the management of research data. See also A.2 (Other technical terms and acronyms) above.

SCAPE: SCAlable Preservation Environments. European research project, financed under FP7.

SeaDataNet: Pan-European infrastructure for ocean & marine data management. European research infrastructure, ENVRIplus partner.

SIOS: Svalbard Integrated Arctic Earth Observing System. European research infrastructure, ENVRIplus partner.

UCPH: University of Copenhagen (Københavns Universitet). Danish university, ENVRIplus participant.

UEDIN: University of Edinburgh. UK university, ENVRIplus participant.

UGOT: University of Gothenburg (Göteborgs Universitet). Swedish university, ENVRIplus participant.

UHEL: University of Helsinki (Helsingin Yliopisto). Finnish university, ENVRIplus participant.

UiT: University of Tromso (Universitetet i Tromsø). Norwegian university, ENVRIplus participant.

UniHB: University of Bremen (Universität Bremen). German university, ENVRIplus participant.

UNILE: University of Salento (Universitá del Salento). Italian university, ENVRIplus participant.

UNITUS: University of Tuscia (Universitá Degli Studi della Tuscia). Italian university, ENVRIplus participant.

USTAN: The University Court of the University of St. Andrews. UK university, ENVRIplus participant.

UvA: University of Amsterdam (Universiteit van Amsterdam). Dutch university, ENVRIplus participant.

A.4. ENVRIplus project-related acronyms & terms

AC: Active Collab (ENVRIplus Project Management System)

BEERi: Board of European Environmental Research Infrastructures - is an internal advisory board representing the needs of environmental Research Infrastructures

CA: Consortium Agreement - Legal contract between the ENVRIplus beneficiaries

DL: Deliverable / Deadline

DoA: Description of Action

DoW: Description of Work

EB: Executive Board - supervisory body for the execution of the Project

EC: European Commission - is the executive body of the European Union responsible for proposing legislation, implementing decisions, upholding the EU treaties and managing the day-to-day business of the EU

EINFRA-1-2014: H2020 Call for e-infrastructures (Managing, preserving and computing with big research data), funding source for ENVRIplus

ENV SWG: the Strategic Working Group on Environment of ESFRI

ENVRI: FP7 project on Implementation of common solutions for a cluster of ESFRI infrastructures in the field of environmental Sciences. Precursor of ENVRIplus.

ENVRIplus: Horizon 2020 project bringing together Environmental and Earth System Research Infrastructures, projects and networks together with technical specialist partners to create a coherent, interdisciplinary and interoperable cluster of Environmental Research Infrastructures across Europe.

ESFRI: the European Strategy Forum on Research Infrastructures

GA: 1) Grant Agreement - Contract between Coordinator and Commission; 2) General Assembly - GA is the ultimate decision-making body of the consortium

H2020: Horizon 2020, European level research funding scheme

I3: Integrated Infrastructures Initiative (I3) combines several activities essential to reinforce research infrastructures and to provide an integrated service at the European level

INFRADEV-4: Sub-call topic of the H2020 INFRADEV call for Implementation and operation of cross-cutting services and solutions for clusters of ESFRI and other relevant research infrastructure initiatives

PM: Person Month

RI: Research Infrastructure. RIs are facilities, resources and related services used by the scientific community to conduct top-level research in their respective fields, ranging from social sciences to astronomy, genomics to nanotechnologies.

VCP: Virtual Community Platform

WP: Work Package

APPENDIX B. AGENDA OF THE "CLOSING THE GAP" WORKSHOP

The one-day workshop took place on October 18, 2017, at the German climate computing centre (DKRZ) in Hamburg, Germany. See **Appendix C** for abstracts of all presentations. The outcomes of the closing discussion are summarized in **Chapter 3.3.1**.

09:00 – 09:15 Welcome and introduction

- *Welcome* - Frank Toussaint, DKRZ
- *Introduction to ENVRIplus and Work Package 6* - Maggie Hellström, ICOS

09:15 – 10:30 Use cases ENVRIplus research infrastructures

- *A single DOI for Argo; a generic approach to making datasets that grow and evolve with time citable on legacy infrastructure* - Justin Buck, National Oceanography Centre
- *Recent developments of the data citation services at WDCC/DKRZ* - Martina Stockhause, DKRZ
- *From data archival to citation through PIDs and DOIs. The GEOFON use case* - Javier Quinteros, GFZ
- *ICOS and data citation* - Alex Vermeulen, ICOS

10:30 – 11:00 Coffee

11:00 – 12:30 PID for non-data research objects

- *Documentation and identification of long term monitoring facilities* - Johannes Peterseil, Umweltbundesamt GmbH
- *Persistent Identification of Instruments* - Markus Stocker, Universität Bremen
- *Linking Environmental Data and Samples* - Kerstin Lehnert, SESAR, Columbia Unversity
- *Application aware digital objects access and distribution using Named Data Networking (NDN)* - Zhiming Zhao, University of Amsterdam (remote presentation)
- *The VAMDC Query Store* - Carlo Maria Zwölf, Paris Observatory

12:30 – 13:15 Lunch

13:15 – 14:30 PID service providers and data publishers

- *RDA perspective: PID Kernel Information and registries within the Data Fabric context* - Tobias Weigel, DKRZ
- *Data Identification and Tracing Services of ePIC* - Ulrich Schwardmann, GWDG, ePIC
- *Supporting data citation on Research Infrastructures using PID-based workflows* - Kristian Garza, DataCite
- *How (and why) to get citations for your data* - Edward van Lanen, Elsevier PANGAEA, Robert Huber, Universität Bremen

14:30 – 15:00 Coffee

15:00 – 16:00 Discussion on future directions

16:00 – 16:05 Close

APPENDIX C. PRESENTATIONS MADE AT THE "CLOSING THE GAP" WORKSHOP

Here we list all presentations made at the workshop, ordered according to the sequence of presentations. For each contribution, a brief abstract is given.

A single DOI for Argo; a generic approach to making datasets that grow and evolve with time citable on legacy infrastructure

Justin Buck, National Oceanography Centre UK & Euro-Argo

The Argo dataset grows and evolves with time and changes in the expectations on the citation of Argo data and traceability of data citations has driven a 5 year effort to make Argo data citable via a single DOI. This has now been implemented by Ifremer on the Argo dataset for the first time using an approach that enables citation for the Argo data at monthly snapshots without requirement for significant enhancement to the Argo data infrastructure. The approach presented is readily applicable to other data infrastructures and enables Argo to partly meet the recommendations of the Research Data Alliance Dynamic Data Citation working group.

Recent developments of the data citation services at WDCC/DKRZ

Martina Stockhause, World Data Centre for Climate & German Climate Computation Centre (DKRZ)

The talk will give a brief overview of the recent developments and future plans of the citation services at WDCC/DKRZ. The currently developed CMIP6 (Coupled Model Intercomparison Project Phase 6) and IPCC AR6 (Intergovernmental Panel on Climate Change Assessment Report 6) data citation services will serve as an example. With its various data-literature, data-data and data-scientist/institution interlinks, the CMIP6/AR6 data citation concept supports tracing of data usage in literature and scientific projects via the Scholix services.

From data archival to citation through PIDs and DOIs. The GEOFON use case.

Javier Quinteros, GFZ Helmholtz Centre Potsdam & GEOFON (EPOS)

A presentation of our current data workflow and our work-in-progress related to PIDs, Data Collections and DOIs.

ICOS and data citation

Alex Vermeulen, Lund University & ICOS

ICOS is a research infrastructure that provides observations of the carbon cycle, targeted at scientific users. A distributed network of more than 120 stations divided over national networks provides high quality and precision measurements from atmosphere, ecosystems and ocean. Curation and data processing is performed by also distributed thematic centres. The ICOS Carbon Portal gives access to all ICOS data and metadata. ICOS data is distributed using a CC4BY license where proper citation is required that users have to accept before they can access the data and that has to be passed on with the data at redistribution. ICOS will provide a specific citation with each data object download. The citation contains a persistent identifier (e.g. DOI) that will link to a landing page with all relevant metadata, including information on the contributors. Data download statistics will be gathered and applied to the records of all contributors by the Carbon Portal.

However when data objects are cited on the web or in literature it is still an open issues how to harvest this use and attribute the contributors.

Documentation and identification of long term monitoring facilities

Johannes Peterseil, Environment Agency Austria & LTER Europe

The proper documentation of observation facilities is a core part of any site based observation network providing sufficient information on the context of the observation. The DEIMS Site and Dataset Registry (https://data.lter-europe.net/deims/) provides a web based catalogue to document observation facilities. DEIMS-SDR is used by LTER Europe, ILTER and beyond. In order to be able to provide a unique identification of these sites work is done to set up a DEOS-ID (Digital Environmental Observation Site Identifier) which could be used across different site catalogues. Within the DEIMS-SDR a prototype implementation will be created linking community based site documentation with the DEOS-ID.

Persistent Identification of Instruments

Markus Stocker, PANGAEA/MARUM & University of Bremen

To interpret a dataset, we need contextual information about the hardware used to generate the data. This talk will introduce to persistent identification of instruments and focus on one aspect: the why and how to involve manufacturers. We will also give an update on plans for a corresponding RDA WG.

Linking Environmental Data and Samples

Kerstin Lehnert, Lamont-Doherty Earth Observatory at Columbia University, USA & IGSN e.V.

Samples and the data generated by their studies represent one of the primary foundations of environmental research and are key to our knowledge of Earth's dynamic systems, its state and evolution. Open, transparent and reproducible science demands samples that are the object of studies and pertinent publications and data are discoverable, accessible, and re- usable, with interoperable metadata in online catalogues. This presentation will provide an overview of best practices for unique and persistent identification of samples, sample registration, sample documentation, and related policies.

Application aware digital objects access and distribution using Named Data Networking (NDN)

Zhiming Zhao, University of Amsterdam

In big data infrastructures, Persistent Identifiers (PIDs) are widely used to identify digital content and research data. A typical example of PIDs is the Digital Object Identifier (DOI). In a data centric application (such as a scientific workflow) it is often required to fetch different data objects from multiple locations. When reproducing a workflow published by community, data objects involved in the workflow often have PIDs. In this project we investigated how to optimize the fetching and sharing of DOI identified objects with Information centric networking paradigm such as Named Data Networking (NDN). In order to achieve that goal, first we presented an approach for integrating PIDs with Named Data Networking (NDN) networks. NDN identifies digital objects with their names and route them also based on their names. In addition, we proposed an approach for optimizing the NDN network's performance using application level knowledge, such as the size, number, and order of the requested objects. We investigated the effect of ordering a group of objects in ascending or descending order according to their sizes before requesting them one by

one. The results showed that the order of the requests can dramatically influence performance of fetching objects from NDN networks.

The VAMDC Query Store

Carlo Maria Zwölf, Paris Observatory & Virtual Atomic and Molecular Data Centre

The Virtual Atomic and Molecular Data Centre (VAMDC) federates ~30 heterogeneous databases, providing a "wrapper layer" that allows to expose the data in a unified way. The talk presented details of the VAMDC infrastructure's technical architecture, highlighting the VAMDC Query Store - an implementation of the RDA Dynamic Data Citation working group recommendation. The Query Store approach allows to both store all necessary information and metadata that is needed to reliably reproduce search queries made by end users, and to make these consistently citable via assigned persistent identifiers. The presentation also addressed a number of encountered issues, including problems for human end users to interpret the XML-formatted QS output, the inability of some end user software clients to properly parse the QS output, and the need for end users to modify their work processes to properly capture provenance information. The VAMDC experiences highlight a number of questions that are of general interest for the science community as a whole: How to best educate end users on how to use persistent identifiers for data search, provenance, citation etc.? Will publishers contribute to costs linked with the storage and digital curation of data? How should credit be assigned and distributed also to non-authors, including curators, data managers, maintainers of the necessary e-infrastructure etc.? *(Summary provided by the workshop organizers.)*

The RDA perspective: PID Kernel Information and registries within the Data Fabric context

Tobias Weigel, German Climate Computation Centre (DKRZ)

This talk will provide a brief update on recent RDA activities concerning the conceptual and use case oriented discussions around PID Kernel Information, their possible definition, sharing and workflow-enablement through registries, relationship with collections and the overall significance of this in the architectural framework that is discussed within the Data Fabric group. The talk will set these activities also in the context of the data citation and tracking challenges relevant for environmental RIs and publishing workflows.

Data Identification and Tracing Services of ePIC

Ulrich Schwardmann, ePIC & Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen

After a short introduction to ePIC this talk presents the services ePIC provides to ensure reliable data identification and data tracing for research data. The granularity of the data presentation is a major topic of the scientific data management in its life cycle from data generation to data reference and citation and the reuse of data. PIDs, data fragments, data collections and registered data types are useful tools to structure this granularity problem.

Supporting data citation on Research Infrastructures using PID-based workflows

Kristian Garza, DataCite

Data citation is a key practice on supporting research data access, sharing and reuse, as well as reproducible scholarship. However, supporting the unveiling the citation links between scholarly

literature and underpinning research data is still a problem. DataCite has been using automatic workflows based on Persistent Identifiers to address this problem. These workflows capture and aggregate citation events that happen in either the Publisher or the Research Infrastructures. These workflows can deliver relationships between people's ORCIDs and research data as well relationships between scholarly literature and the underpinning research data. So far these workflows have been successful to connect people with ORCIDs to their data and DataCite can send that information back to Publishers and Research Infrastructures. We argue that data publication workflows represent the path forward to address the problems the linking and retrieving information about data identification and data citation.

How (and why) to get citations for your data

Edward van Lanen, Elsevier

There are many reasons to make data available for others to use, including improving quality of research by increasing transparency and reproducibility - but also a wish to increase discoverability of research to enhance the possibility for receiving credit, as well as mandates from funders. Currently about half of polled researchers say they share data, with a majority doing so via supplemental materials attached to publications. Many use personal, institutional or project websites to link to data, with much fewer making use of either general or discipline-specific repositories. Posting data mainly as supplementary materials, however, has many issues, including no or poor searchability, inability to cite (and receive credit for) the data, and poor coverage of data not directly connected to articles. Elsevier is now launching a new service, Data in Brief, to provide a way for researchers to more easily share and reuse each other's datasets by publishing them as data articles. *(Summary provided by the workshop organizers.)*

PANGAEA - Data Publisher for Earth & Environmental Science

Robert Huber, University of Bremen & PANGAEA/Marum

The ISCU World Data Centre PANGAEA is an information system for acquisition, processing, long term storage, and publication of geo-referenced data related to earth science fields. Storing more than 350.000 data sets from all fields of geosciences it belongs to the largest archives for observational earth science data. Standard conform interfaces (ISO, OGC, W3C, OAI) enable access from a variety of data and information portals, among them the search engine of PANGAEA itself ((www.pangaea.de) and e.g. GBIF and GEOSS. All data sets in PANGAEA are citable, fully documented, and can be referenced via persistent identifiers (Digital Object Identifier - DOI) - a premise for data publication. Together with other ICSU World Data Centres (www.icsu-wds.org) and the Technical Information Library in Germany (TIB) PANGAEA had a share in the implementation of a DOI based registry for scientific data, which by now is supported by a worldwide consortium of libraries (www.datacite.org). A further milestone was building up strong co-operations with science publishers as Elsevier, Springer, Wiley, AGU, Nature and others. A common web service allows to reference supplementary data in PANGAEA directly from an articles abstract page (e.g. Science Direct). The next step with science publishers is to further integrate the editorial process for the publication of supplementary data with the publication procedures on the journal side. PANGAEA is operated as a joint long term facility by MARUM at the University Bremen and the Alfred Wegener Institute for Polar and Marine Research (AWI). More than 80% of the funding results from project data management and the implementation of spatial data infrastructures (more than 160 International to national projects) since the last 15 years - www.pangaea.de/projects.

APPENDIX D. SURVEY PARTICIPANTS

The following is a list of all the organisations who participated in the survey. We would like to express our sincere thank you to all of them for their engagement and willingness to share their views with us.

TABLE D-1. THE NAME, RESPONDENT CATEGORY AND BRIEF DESCRIPTION OF ALL THE ORGANISATIONS WHO PARTICIPATED IN THE QUESTIONNAIRE-BASED SURVEY.

Name and URL of organisation	Category	Description
Clarivate Analytics https://clarivate.com	Data usage indexers	Company delivering products for market, research and technology analysis. Examples of products are Web of Science, MarkMonitor, TechStreet etc.
Copernicus.org https://www.copernicus.org/	Publisher & publisher association	Publisher of Open Access journals
CrossRef https://www.crossref.org/	PID provider	Membership based not-for-profit organisation delivering PID and metadata services
DataCite https://www.datacite.org/	PID provider	Not-for-profit organisation delivering PID and metadata services
EBSCO Information Services http://www.ebsco.com	Publisher & publisher association	Major publisher delivering databases and journals in multiple sectors and areas.
Elsevier https://www.elsevier.com/	Publisher & publisher association	Major publisher delivering databases and journals in multiple sectors and areas.
ePIC http://www.pidconsortium.eu/	PID provider	Consortium for PID services based on the Handle system.
EUDAT https://eudat.eu/	PID provider	European infrastructure delivering services for PIDs and other research data related services and systems.
Hindawi https://www.hindawi.com/	Publisher & publisher association	Publisher of Open Access journals.
NISO (National Information Standards Organization) https://www.niso.org/	Other	NISO is a non-profit association accredited by the American National Standards Institute (ANSI), developing standards to manage information in today's continually changing digital environment.
NOAA (National Ocean and Atmospheric Administration) http://www.noaa.gov/	Other	National agency for studying and predicting changes in climate, weather, oceans, and coasts.
OASPA https://oaspa.org/	Publisher & publisher association	Association for Open Access publishers.

Name and URL of organisation	Category	Description
Open Citations http://opencitations.net/	Data usage indexers	The main work of OpenCitations is the creation and current expansion of the Open Citations Corpus (OCC), an open repository of scholarly citation data.
ORNL DAAC https://earthdata.nasa.gov/about/daacs/daac-ornl	PID provider	ORNL DAAC provides data and information relevant to biogeochemical dynamics, ecological data, and environmental processes, critical for understanding the dynamics relating the biological, geological, and chemical components of Earth's environment.
PANGAEA https://www.pangaea.de/	Publisher & publisher association	Data publisher for earth & environmental Science.
re3data.org http://www.re3data.org/	Other	re3data offers detailed information on more than 2,000 research data repositories.
SESAR http://www.geosamples.org/	PID provider	SESAR operates a registry that distributes the International Geo Sample Number IGSN. SESAR provides access to the sample catalogue via the Global Sample Search.
STM https://www.stm-assoc.org/	Publisher & publisher association	STM is a membership based organisation for academic and professional publishers.

APPENDIX E. SELECTED EXCERPTS FROM THE QUESTIONNAIRE RESPONSES

In order to complement the discussion in **Chapter 4.3.2**, we present here a subset of responses to our survey questionnaire. For each excerpt, the category of the respondent (publisher, PID service provider, data usage indexer or other) is indicated. Note that in order to protect the anonymity of the respondents, and to harmonize the style, the excerpts have been slightly modified.

Q1: The concept "persistent" in persistent identifiers, what does that mean to you and the services in your organization?

- "That digital objects can always be found via the PID and that links to it will not break." (publisher)
- "Developing the social and technical infrastructure that will enable content to be perpetually identified." (PID service provider)

Q4: What is your opinion on PID based references pointing to samples, instruments and stations in scientific articles, e.g. PIDs to non-data objects? Would it be feasible to support PID services to these references to non-data objects?

- "Having a PID for instruments and stations will improve understanding and reproduction of data." (PID service provider)
- "PIDs should be assigned to such entities, resolving to a landing page describing the entity." (other)

Q6: What is your opinion on allowing bibliographic references to be made to dataset fragments or subsets (i.e. by appending pointer information to the PID of the dataset)? Do your services support pointers to subsets in bibliographic references?

- "It depends on the community and the need for a citation. If a fragment needs to be cited then it is entirely appropriate." (PID service provider)
- "As a basic principle, one identifies the thing at the most granular level that it is necessary and practical for business purposes to identify a thing. The use of suffixes in identifiers in essence is simply a different identifier. Rather than building semantics into the string of the identifier, simply use a different ID and link the two objects using the relevant metadata." (other)

Q7: What is your opinion about using PIDs for data collections (i.e. collections of several datasets)? Do your services support PIDs for data collections?

- "In some cases it is useful, but we need to ensure that credit is given to authors of the individual objects in the data collection." (PID service provider)
- "We encourage using PID's for data collections." (publisher)

Q8a: Do your services support bibliographic references for dynamic data sets?

- "There is a valid use case for PIDs for dynamic data. It is incumbent on the data producer or provider to provide necessary context and metadata so it is understood by the user." (other)
- "We have DOIs for datasets that are streaming in from sensors. We also have a versioning system where datasets can be updated or revised and keep the same DOI." (PID service provider)

Q8b: What is your opinion on assigning PIDs to search queries rather than assigning PIDs to the results of a query?

- "This would require rigorous versioning of databases to ensure that saved queries produce the same results, but it could be useful to avoid storing large datasets and subsets generated by dynamic searches." (PID service provider)
- "This is not a good idea. There are an infinite number of possible queries, most of which are rather than persistent, and PIDs should only be used to point to durable objects." (other)

Q9a: In relation to persistence of PIDs, what is your opinion on the "sustainability" of your products and services?

- "Our PID's are persistent and sustainable because they are founded on a solid business model." (publisher)
- "Our service is fairly sustainable as we have a robust membership structure and have been fully integrated into scholarly publishing workflows." (PID service provider)

Q9b: What time frame would constitute "sustainable" for your services?

- "Our business will hopefully persist for decades, while our content should persist for centuries." (publisher)
- "At least 10 years, ideally longer. " (data usage indexer)