Identification and Citation of Digital Research Resources

Margareta Hellström[1], Maria Johnsson[2], and Alex Vermeulen[3(✉)]

[1] Department of Physical Geography and Ecosystem Science, Lund University, Sölvegatan 12, 22362 Lund, Sweden
margareta.hellstrom@nateko.lu.se
[2] Department of Scholarly Communication, Lund University Library, Box 3, 221 00 Lund, Sweden
maria.johnsson@ub.lu.se
[3] ICOS ERIC - Carbon Portal, Sölvegatan 12, 22362 Lund, Sweden
alex.vermeulen@icos-ri.eu

Abstract. Environmental research infrastructures are often built on a large number of distributed observational or experimental sites, run by hundreds of scientists and technicians, financially supported and administrated by a large number of institutions. It becomes very important to acknowledge the data sources and their providers. There is also a strong need for common data citation tracking systems that allow data providers to identify downstream usage of their data so as to demonstrate their importance and show the impact to stakeholders and the public. This chapter highlights identification and citation in environmental RIs, reviews available technologies and develops common services for these operations. This chapter presents a suggested common system design for Identification and Citation, as well as an outline for negotiations and discussions with publishers and other actors in the scholarly data management and curation world.

Keywords: Identification · Citation · Persistent identifiers

1 Introduction

To perform data intensive sciences one often requires data that are managed by different institutions. Observations and measurements from infrastructures in environmental and earth sciences are a particularly strong example of this multitude of sources. The result of the scientific analysis based on this data depends heavily on the access to high quality data and the proper citing of those data sources or data sets when publishing the final results becomes an important practice for acknowledging the original data providers and for keeping the study reproducible.

Environmental research infrastructures are often construed from a large number of distributed observational or experimental sites, run by hundreds of scientists and technicians, financially supported and managed by a large number of institutions. If these data are shared under an open access policy this becomes another important reason to

Z. Zhao and M. Hellström (Eds.): Towards Interoperable Research Infrastructures for Environmental and Earth Sciences, LNCS 12003, pp. 162–175, 2020.
https://doi.org/10.1007/978-3-030-52829-4_9

acknowledge the data sources and their providers. A data citation tracking system that allows the data providers to identify downstream usage of their data and assess the impact of their data to stakeholders and the public is then a strong requirement.

Furthermore, a common policy model is needed for persistent identifiers for publishing and citing data. Moreover, the services for assigning and handling identifiers and for retrieving data content based on identifiers will also have to be provided.

In this chapter we will discuss the building blocks to fulfil theses needs, building on existing approaches and current activities undertaken by ENVRI partners, and—if needed—synchronise with developments that arise from up-coming studies and projects from both service providers (e.g. ePIC, DataCite and EUDAT) and initiatives based in research organizations (e.g. THOR and OpenAIRE). The work described has been operated in close cooperation with existing initiatives (e.g. Research Data Alliance and ICSU WDS) and will elaborate a common data citation solution for the involved RIs.

This chapter presents a strategy developed during the ENVRIplus project to negotiate with external organisations. The content is mainly based on the public deliverable D6.1, D6.2 and D6.3 of the ENVRIplus project[1].

2 Background

2.1 Identification

A number of approaches have been applied to solve the question of how to unambiguously identify digital research data objects [1]. Traditionally, researchers have relied on their own internal identifier systems, such as encoding identification information into filenames and file catalogue structures, but this is neither comprehensible to others, nor sustainable over time and space [2]. Instead, data object identifiers should be unique "labels", registered in a central registry database that contains relevant basic metadata about the object, including a pointer to the location where the object can be found as well as basic information about the object itself. Exactly which metadata should be stored in the identifier registry, and in which format, is a topic under discussion, see e.g. [3]. Many environmental observational datasets pose a special challenge in that they are not reproducible, which means that also fixity information (checksums or even "content fingerprints") should be tied to the identifier [4].

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

In [1], the authors provide a comprehensive summary of the pros and cons of different identifier schemes, and also assess nine persistent identifier technologies and systems. Based on a combination of technical value, user value and archive value, DOIs (Digital Object Identifiers provided by DataCite) scored highest for overall functionality, followed by general handles (as provided by e.g. CNRI and DONA) and ARKs (Archive

[1] EU H2020 ENVRIplus www.envriplus.eu.

Resource Keys). DOIs have the advantage of being well-known to the scientific community via their use for scholarly publications, and this has contributed to their successful application to e.g. geoscience datasets over the last decade [5]. General Handle PIDs have up to now mostly been used to enable referencing of data objects in the pre-publication steps [6] of the research data life cycle (illustrated in Fig. 1). They could however in principle equally well be applied to finalised "publishable" data.

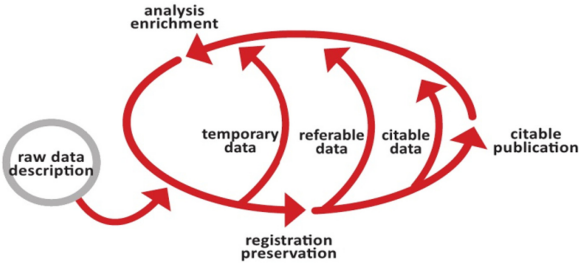

Fig. 1. The research data life cycle. Data intensive research is highly collaborative. Allocating persistent identifiers to data objects supports (re-)use and sharing of data also in early stages of the research life cycle [6].

Persistent identifiers systems are also available for research-related resources other than digital data & metadata, articles and reports—it is now possible to register many other objects, including physical samples (IGSN), software, workflow processing methods—and of course also people and organisations (ORCID, ISNI). In the expanding "open data world", PIDs are an essential tool for establishing clear links between all entities involved in or connected with any given research project [7].

2.2 Citation

The FORCE11 Joint Declaration of Data Citation Principles (JDDCP) [8] states that in analogy to articles, reports and other written scholarly work, also data should be considered as legitimate, citable products of research. (There is however currently an ongoing discussion as to whether datasets are truly "published" if they haven't undergone a standardised quality control or peer-review, see e.g. [9].) Thus, any claims in scholarly literature that rely on data must include a corresponding citation, giving credit and legal attribution to the data producers, as well as facilitating the identification of, access to and verification of the used data (subsets). A generic workflow for data citation is presented in Fig. 2. The workflow consists of a citation from a document to a dataset, a landing page in the repository where the dataset is stored, and the dataset itself.

Data citation methods must be flexible, which implies some variability in standards and practices across different scientific communities [8]. However, to support interoperability and facilitate interpretation, the citation should preferably contain a number of metadata elements that make the dataset discoverable, including author, title, publisher, publication date, resource type, edition, version, feature name and location. Especially important, the data citation should include a persistent method of identification that is

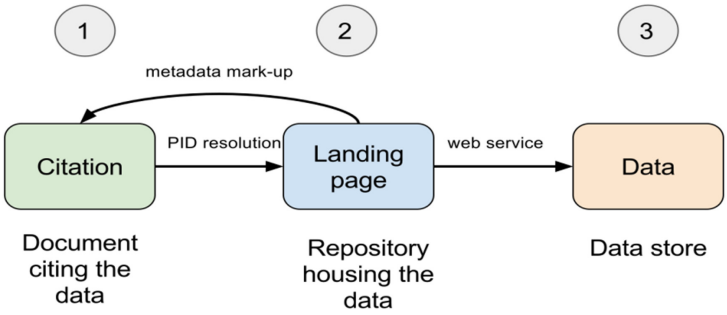

Fig. 2. A generic data citation workflow. (http://force11.github.io/data-citation-primer/)

globally unique and contains the resource location as well as (links to) all other pertinent information that makes it human and machine actionable. In some (sensitive) cases, it may also be desirable to add fixity information such as a checksum or even a "content fingerprint" in the actual citation text [4].

Finding standards for citing subsets of potentially very large and complex datasets poses a special problem, as outlined by [10], as e.g. granularity, formats and parameter names can differ widely across disciplines. Another very important issue concerns how to unambiguously refer to the state and contents of a dynamic dataset that may be variable with time, e.g. because new data are being added (open-ended time series) or corrections introduced (applying new calibrations or evaluation algorithms) [11].

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

A number of surveys have indicated that the perceived lack of proper attribution of data is a major reason for the hesitancy felt by many researchers to share their data openly [4, 12, 13]. 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.

3 Components of PID Systems

3.1 Common PID Types: The Persistent Identifier Zoo

In this section, we present an overview of seven of the most commonly used persistent identifier types. The underlying study was performed in the summer of 2016 by Huber and co-workers, and the numbers and statistics represent the status of the re3data.org registry[2] at that time.

The Handle System (HS)

Arguably the biggest impact in the field of persistent identification of digital research resources was achieved by the Handle System [14]. The Handle System (HS) describes a minimal set of requirements for an infrastructure for the identification of objects in

[2] http://www.re3data.org/.

a digital infrastructure and how the identity of an object can be related to its location. The system is agnostic to the contents of the objects, keeping it open for interoperability with future applications. The HS separates the identifier from the resolving mechanism, making it independent of HTTP and DNS but in practice, the system is mostly leveraged using a HTTP proxy that allows the use of a RESTful API and URLified handles. The HS supplies a stable, distributed platform for the resolution of identifiers to URLs, including methods more sophisticated than HTTP redirects like template handles and embedded metadata.

In the sample of 1381 repositories listed in the re3data repository at the time of the study, the HS is used by 102 repositories. Handle is mainly used by institutional repositories, which might be linked to the role of Handle as an identifier in repository software like DSpace[3].

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

Digital Object Identifier (DOI)

Looking at the 475 repositories using any kind of PID system, the most commonly implemented identifier type was the digital object identifier (DOI). DOIs, which were introduced in 1998 by the International DOI Foundation (see http://www.doi.org/), were used by 275 out of those 475 repositories, meaning that the use of the DOI eclipsed all other persistent identifiers. The use of DOI persistent identification of data initiated by a project funded by the German research foundation in 2003 [5]. DOI were chosen because of their already established part in the scholarly publication infrastructure.

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

"Cool" Uniform Resource Identifiers (CoolURIs)

Compared to the strict criteria of Nestor [15] and other related efforts, "cool" (meaning unchanging or static) Uniform Resource Identifiers (CoolURIs) somehow represent an anarchic view on identifiers. Similar to URN, the idea of CoolURIs goes back to early ideas about identification and location of objects on the web. The idea of CoolURI [16] is fundamental for the Semantic Web. It is based on Uniform Resource Identifiers (URIs) which, by proclamation, will not change. They make use of standard HTTP functionalities, in particular content negotiation[4], to enable the URI to be resolved to different representations (RDF, HTML) of the same object. CoolURIs allow webmasters to maintain the persistence of their resource identifiers, the URIs, with a minimum of effort and without a centralised PID system.

[3] http://www.dspace.org/.

[4] https://en.wikipedia.org/wiki/Content_negotiation.

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

The CoolURI concept relies on HTTP as resolving mechanism and assumes that the HTTP protocol will be around for a long time.

3.2 Identifiers for Non-data Entities

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

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

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

Identifiers for Instrumentation and Sensors
By assigning unique and persistent identifiers to sensors and other instrumentation, and using these PIDs consistently in both cataloguing and curation, researchers can simplify the management and collection of observation metadata records, and facilitate property lookup and provenance tracing throughout all steps of the research data processing cycle.

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

Identifiers for Data Content Types
In order to facilitate (re-)use of datasets, especially in the context of machine-actionable workflows, it is useful to make use of persistently identified Data Type definitions. These should include a basic description of the characteristics of a given data or variable, but can also contain information on which software should be used to process it further. See e.g. the recommendations of the RDA Data Type Registries working group (https://www.rd-alliance.org/group/data-type-registries-wg/outcomes/data-type-registries).

Identifiers for Software
GitHub[5] and similar software repositories support versioning, and as such allow the code author to link directly by URL to a specific code package or file. In GitHub, objects can themselves be linked to dataset DOIs, so there are possibilities of cross-referencing. However, at the moment it is not yet possible to provide a DOI or any other PID to software codes or packages in GitHub. Notably, the German Climate Computing Centre DKRZ (see Sect. 7.1.2) is about to apply for national project funding to offer sustainable production and long term storage of scientific software. This will account for versioning and include the use of persistent IDENTIFIERS. See also Sect. 8.3.1.

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

4 Identification and Citation in Practice—Recommendations to RIs

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

4.1 Identification Best Practices for RIs

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

[5] https://github.com/.

- data objects (files, databases etc.)
- metadata objects
- articles, reports and other documents related to the data
- people, including everyone involved in the data production chain
- organizations (agencies, institutes, and RIs themselves) involved in the data production chain
- sensors and sensor platforms, measurement stations, cruises, measurement campaigns
- physical samples.

In addition, comprehensive use of PIDs should be considered for

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

4.2 Citation Best Practices for RIs

RIs should strive to follw the following recommendations for data citation, based on the review of data citation best practices and recommendations from relevant organisations including [4, 8, 17]:

Technical aspects:

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

Citation practices:

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

5 Cases in ENVRI

5.1 Development of a Citation and Usage Tracking System for Greenhouse Gas

ICOS, Integrated Carbon Observation System, is a pan-European research infrastructure with a mission to provide standardised, long term, high precision and high-quality observations on the carbon cycle and Green House Gas (GHG) budgets and their perturbations. The ICOS observing network consists of over 140 observation stations, each related to one or more of the three domains Atmosphere, Ecosystem and Ocean, and operated by its (currently 14) member countries. The collected data is processed and quality controlled at Thematic Centres (one for each domain), before being openly distributed via the ICOS data centre named Carbon Portal (CP).

The ICOS Carbon Portal is designed to manage and/or distribute data objects of a number of different categories. All data objects are assigned a Handle System-based persistent identifier at the time of ingestion into the Carbon Portal repository. The PID of these individual data objects can be resolved via e.g. the handle.net resolver service, which redirects to the object's landing page hosted by the Carbon Portal. The landing page lists the most relevant metadata of the object, including a direct link to access to the data object.

Objects can be registered with DataCite, either as single objects or as collections of data objects. This means that they are also assigned a DOI and that associated metadata are stored in the DataCite catalogue following the DataCite Metadata Schema[6]. This allows the data objects to be found also through searches on the DataCite portal, and also provides full integration with the Citation Formatter service from Crossref and DataCite.

Resolving the DOI, e.g. via the resolving services of handle.net and the DOI Foundation, results in a redirect to the landing page of the object or object collection, hosted by the Carbon Portal. The DataCite DOI identifiers have the form "10.18160/<suffix>", where "10.18160" is the ICOS-specific DataCite prefix, and <suffix> is a globally unique string. (The strings used for DOIs are also computed starting from the data object hash sum, but are designed to be shorter and easier to read for humans. As in the case of the Handle PIDs mentioned above, additional tests for uniqueness are of course performed before submitting the DOI registration request.)

Any updated versions of a given object, for example dynamic, growing time series or corrected data sets, are assigned a completely new PID. In order to provide unbroken provenance chains, the metadata record of the old version is updated with a link to the superseding object, and vice versa. This strategy is applied to data objects of all types described above.

The Carbon Portal provides a landing page (as shown in Fig. 3) for any digital object described in the metadata store – including data sets, but also stations, data type specifications and concepts. All landing pages are created dynamically, i.e. at the moment that their URL is accessed. This means that the displayed information always reflects the current, most up-to-date information. The format, and what information is shown, will vary between the type of objects, with the richest content provided for data objects.

[6] https://schema.datacite.org/.

Fig. 3. Screenshot of the ICOS landing page for the data object ICOS_ATC_L2_L2-2018.1 - Collection of ICOS ATC L2 data objects (release 2018.1). The page exposes basic metadata about the data set itself, its contents, and a recommended citation string – the latter includes the data set's DataCite DOI number, https://doi.org/10.18160/RHKC-VP22.

ICOS data is distributed using a Creative Commons Attribution 4.0 International (CC BY 4.0) license[7], which users have to accept before they can access the data. The CC BY license requires end users to give appropriate credit (i.e. citing data when it is used), provide a link to the license, and indicate if changes were made when re-distributing the data. When citing ICOS data, at a bare minimum the data object's persistent digital identifier should be given in a machine-actionable form (as a HTTP URL, for example http://hdl.handle.net/11676/6PrNhZelwXKHLqO41QRsbheu or https://doi.org/10.18160/RHKC-VP22); this minimal form is sufficient for inclusion in provenance records, but for use in scientific literature much more information (contributors, data set name etc.) is of course required.

5.2 Facilitating Quantitatively Correct Data Usage Accounting

In a world of open and free data sharing, it is often necessary to document the use of data products and give this as a quantitative merit to data producers and providers. Since all entities involved in the data production chain face the challenge of having to find sources of continued funding for their efforts while "selling" their data is not an option. They need to justify to funding agencies and users the relevance of their observations and contributions to data production.

An existing analogy to such a use-based merit system are scientific journal publications, where authors receive merit based on the number of "uses" of the article, i.e. based

[7] https://creativecommons.org/licenses/by/4.0/.

on the number of citations. Journals are selected by authors and institutions based on the aggregation of those citation scores in recent years, i.e. how visible the result becomes by using a given publication channel. However, aggregating scores, i.e., citation numbers accumulated across repositories, may be difficult to compose if data is stored at different granularities in the different archives.

By analogy to scientific articles, persistent identification of data by Digital Object Identifiers (DOIs) would be a crucial element of such a service for quantitative accounting of data use. However, at least 4 challenges exist:

- **DOI granularity:** This would make usage numbers based on a fine granularity biased in comparison to data identified with a coarser granularity.
- **Data collections:** DOIs can refer to a user defined collection of other datasets, which themselves may be identified by DOIs. The data collection approach makes data very convenient to cite. However, the contribution of different data producers to such a collection can vary significantly.
- **Accounting mechanism:** Indexing agencies will or have been setting up services for counting of use events involving (DOI) identified data. From here, services need to be implemented that break down data use events into the contribution of single data producers, and with a fixed granularity allowing comparisons between data producers.
- **Nature of data use events:** Scientific data can be used in many different ways, e.g. illustration for outreach purposes, trend analysis, constraining models of environmental processes, event analysis, just to mention a few, and data can be accessed once or multiple times for the same use case event. A list of data use types counted towards use accounting, including weighting factors, would typically be agreed on and continuously updated by a cross-domain working group consisting of experts on data production, data management, and data indexing.

In order to meet the challenges, and to work towards implementing accounting services for data use, the project team defined the following tasks in the early project phase:

Data Identification with Homogeneous Granularity in Primary Archives
These DOIs, in this context called primary DOIs, would be used as reference for setting up data use accounting. The ambition in the early project phase was to achieve homogeneous granularity, and thus comparable data use metrics, across repositories and RIs. During the implementation, it turned out that the goal of achieving homogeneous granularity of primary data identifiers across atmospheric RIs was too ambitious. Data products are simply too different in nature among repositories, sometimes even within a single RI. However, primary data identifier granularity should be homogeneous at least within one repository, preferably comparable also among repositories of a single RI.

Transparent Data Accounting When Using Data Collections
When identifying data in larger studies, e.g. global climatology of atmospheric parameters, using primary DOIs, requires quoting hundreds of DOIs, which would be rather

inconvenient. The DOI specification provides for coining DOIs for user specified data collections, which are ideally suited to identify data used in larger studies. Ideally, the references would also include further provenance information in order to identify and acknowledge contributors to the data product used.

By interacting with the relevant RDA working group on research data collections[8], this work resulted in a fully finished recommendation for issuing and handling persistent identifiers for data collections that meet the requirement of referencing back to the primary identifiers of the data contained in the collection[9].

Performing Correct Accounting of Data Use

For scientific publications, accounting of use is performed by the indexing agencies. If they offer a similar service for data, it needs to be assured that references to collection identifiers are resolved to the primary identifiers to ensure correct accounting of data use. The task involves a dialogue with the indexing agencies to implement this policy. A dialogue with DataCite as an indexing agency collecting use events involving data DOIs revealed that indexing agencies show little willingness to resolve references to primary identifiers contained in collection DOIs when accounting for data use. From the indexing agencies perspective, this approach makes sense due to the issue of heterogeneous granularity of primary data identifiers across or even within domains. As a result, the task needed to be modified. The service of calculating metrics for data use is moved from the indexing agency to the primary data repository. Based on its own primary DOIs with homogeneous granularity, the primary repository can access data use events stored at the indexing agency, resolve references in collection DOIs, and thus calculate data use metrics comparable across the repository. A prerequisite for this approach would be machine-to-machine access to the indexing agencies data holdings by the primary data archives. A dialogue about this is ongoing and needs to be continued in the near future.

6 Conclusion

This chapter discussed the basic concepts of the data identification and citation, and related standards and best practices. The chapter presented two cases studied in the ENVRIplus project.

Acknowledgements. This work was supported by the European Union's Horizon 2020 research and innovation programme via the ENVRIplus project under grant agreement No 654182.

[8] https://www.rd-alliance.org/groups/research-data-collections-wg.html.

[9] https://github.com/RDACollectionsWG/specification/blob/master/Recommendation%20package/rda-collections-recommendation.pdf.

References

1. Duerr, R.E., et al.: On the utility of identification schemes for digital Earth science data: an assessment and recommendations. Earth Sci. Inform. **4**, 139–160 (2011). https://doi.org/10.1007/s12145-011-0083-6
2. Stehouwer, H., Wittenburg, P.: RDA Europe Analysis Programme: Survey of EU Data Architectures, Deliverable D2.5 from the RDA Europe project (FP7-INFRASTRUCTURES-2012-1) (2015). https://www.rd-alliance.org/data-architecture-survey-report.html. Accessed 16 May 2020
3. Weigel, T., DiLauro, T., Zastrow, T.: PID Information Types WG final deliverable (2015). http://dx.doi.org/10.15497/FDAA09D5-5ED0-403D-B97A-2675E1EBE786
4. Socha, Y.M. (ed.): Out of cite, out of mind: the current state of practice, policy, and technology for the citation of data. Data Sci. J. **12** (2013). Available at https://www.jstage.jst.go.jp/article/dsj/12/0/12_OSOM13-043/_pdf. Accessed 30 Jan 2017
5. Klump, J, Huber, R., Diepenbroek, M.: DOI for geoscience data - how early practices shape present perceptions. Earth Sci. Inform. **9** (2015). https://doi.org/10.1007/s12145-015-0231-5
6. Schwardmann, U.: Epic - Persistent Identifiers For Eresearch (2015). https://doi.org/10.5281/ZENODO.31785
7. Dodds, L., Phillips, G., Hapuarachchi, T., Bailey, B., Fletcher, A.: Creating Value with Identifiers in an Open Data World. Report from Open Data Institute and Thomson Reuters (2014). http://innovation.thomsonreuters.com/content/dam/opeweb/documents/pdf/corporate/Reports/creating-value-with-identifiers-in-an-open-data-world.pdf
8. Data citation synthesis group: joint declaration of data citation principles. In: Martone M. (ed.) FORCE11, San Diego, CA (2014). https://www.force11.org/group/joint-declaration-datacitation-principles-final. Accessed 30 Dec 2016
9. Parsons, M.A., Duerr, R., Minster, J.-B.: Data citation and peer review. EOS Trans. AGU **91**, 297 (2010). https://doi.org/10.1029/2010EO340001
10. Huber, R., Asmi, A., Buck, J., Lucas, J.M.D., Diepenbroek, M., Michelini, A.: Participants Of The Joint COOPEUS/ENVRI/EUDAT PID Workshop: Data citation and digital identifiers for time series data/environmental research infrastructures (2015). http://figshare.com/articles/Data_citation_and_digital_identifiers_for_time_series_data_environmental_research_infrastructures/1285728, https://doi.org/10.6084/M9.FIGSHARE.1285728
11. Rauber, A., Asmi, A., van Uytvanck, D., Proell, S.: Data Citation of Evolving Data: Recommendations of the Working Group on Data Citation (WGDC) (2015). https://zenodo.org/record/1406002
12. Uhlir, P.F.: Rapporteur, For Attribution—Developing Data Attribution and Citation Practices and Standards: Summary of an International Workshop (August 2011), National Research Council (2012). http://www.nap.edu/openbook.php?record_id=13564. Accessed 16 May 2020
13. Gallagher, J., Orcutt, J., Simpson, P., Wright, D., Pearlman, J., Raymond, L.: Facilitating open exchange of data and information. Earth Sci. Inform. **8**, 721–739 (2015). https://doi.org/10.1007/s12145-014-0202-2
14. Kahn, R., Wilensky, R.: A framework for distributed digital object services. Int. J. Digit. Libr. **6**, 115–123 (2006). https://doi.org/10.1007/s00799-005-0128-x
15. Bütikofer, N.: Catalogue of criteria for assessing the trustworthiness of PI systems, nestorMaterialien, Niedersächsische Staats und Universitätsbibliothek Göttingen, Göttingen, Germany. http://nbn-resolving.de/urn:nbn:de:0008-20080710227. Accessed 16 May 2020

16. Berners-Lee, T.: Cool URIs don't change, World Wide Web Consortium (W3C), Cambridge, MA. https://www.w3.org/Provider/Style/URI.html. Accessed 30 Jan 2017
17. Fenner, M.: A data citation roadmap for scholarly data repositories. Sci. Commun. Educ. (2016). https://doi.org/10.1101/097196