Identification and citation

Introduction defining context and scope

General comments

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

For these reasons, updating, or implementing totally new, technology alone does not improve "*usage performance*"[1], as the behaviour of the "*de signated scientific community*" will influence the discoverability and ease of reuse of research data. Scientific traditions and previous investments into soft- or hardware can lead to large time constants for change. Adopting new database technology quickly could, on paper, provide large benefits (to the data providers) like lower costs and easier administration and curation, but may *de facto* be unacceptably lowering overall productivity for significant parts of the user community over a long period of time while the transition is achieved.

Unequivocal identification of resources and objects underlies all aspects of today's research data management. The ability to assign persistent and unique identifiers (PIDs) to digital objects and resources, and to simultaneously store specific metadata (url, originator, type, date, size, checksum, etc.) in the PID registry database, provides an indispensable tool towards ensuring reproducibility of research [Duerr 2011], [Stehouwer 2014], [Almas 2015]. Not only do PIDs enable us to make precise references in reports and literature, but it also facilitates recording of object provenance including explicit relationships between connected objects (data and metadata; parent and child; predecessor and successor), as well as unambiguous descriptions of all aspects and components of workflows [Moreau 2008], [Tilmes 2010]. A pervasive adoption of persistent identifiers in research is expected to contribute significantly to scientific reproducibility 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].

---

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

Background - Identification

A number of approaches have been applied to solve the questions of how to unambiguously identify digital research data objects [Duerr 2011]. Traditionally, researchers have relied on their own internal identifier systems, such as encoding identification information into filenames and file catalogue structures, but this is neither comprehensible to others, nor sustainable over time and space [Stehouwer 2014]. Instead, data object identifiers should be unique "labels", registered in a central 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 registered, and in which formats, is a topic under discussion, see e.g., [Weigel 2015].) 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].

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

Persistent identifiers systems are also available for other research-related resources 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 (Dobbs 2014).

Background - Citation

The FORCE11 Data Citation Principles [Martone 2014] state that in analogy to articles, reports and other written scholarly work, also data should be considered as legitimate, citable products of research. (Although there is currently a discussion as to whether data sets are truly "published" if they haven't undergone a standardised quality control or peer-review, see e.g., [Parsons 2010].) Thus, any claims in scholarly literature that rely on data must include a corresponding citation, giving credit and legal attribution to the data producers, as well as facilitating the identification of, access to and verification of the used data (subsets).

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

Finding standards for citing subsets of potentially very large and complex 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.

Finally, a number of surveys have indicated that the perceived lack of proper attribution of data is a major reason for the hesitancy felt by many researchers to share their data openly [Uhlir 2012], [Socha 2013], [Gallagher 2015]. This attitude also extends to allowing their data to be incorporated into larger data collections, as it is often not possible to perform micro-attribution – i.e., to trace back the provenance of an extracted subset (that was actually used in an analysis) to the individual provider – through the currently used data citation practices.

Change history and amendment procedure

The review of this topic was organised by Margareta Hellström in consultation with the following volunteers: @AlexVermeulen, @HarryLankreijer, @AriAsmi. The major steps in the change history are recorded in the table below. For further details of the complete procedure see item 4 on the Getting Started page.

Date	Name	Institution	Nature of the information added / changed
2016-05-25	Margareta Hellstrom	ULUND & ICOS	Updated page with content from D5.1Candidate2.docx (from 2016-05-24)

Sources of information used

- Web sites of
  - CNRI (Corporation for National Research Initiatives) https://www.cnri.reston.va.us/
  - CrossRef http://crossref.org/
  - DataCite https://www.datacite.org/
  - DataONE https://www.dataone.org/
  - *DDI (Data Documentation Initiative) Alliance* http://www.ddialliance.org/
  - *DONA (Digital Object Numbering Authority)* https://dona.net/
  - ePIC (European Persistent Identifier Consortium) http://www.pidconsortium.eu/
  - EUDAT http://eudat.eu/
  - *euroCRIS* http://eurocris.org/ - responsible for CERIF development
  - ICSU-CODATA (ICSU Committee on Data for Science and Technology) http://www.codata.org/
  - *IGSN (International Geo Sample Number)* http://www.igsn.org/
  - *ISNI (International Standard Name Identifier)* http://www.isni.org/
  - *MDC (Making Data Count project)* - http://mdc.lagotto.io
  - OKNF (Open Knowledge Foundation) https://okfn.org/
  - OpenAIRE https://www.openaire.eu/
  - ORCID http://orcid.org/
  - PANGAEA http://www.pangaea.de/
  - RDA (Research Data Alliance) https://rd-alliance.org/ and the web pages of its many active interest and working groups, including:
    - Bibliometrics Working Group (active)
    - Data Citation Working Group (finished)
    - Data Fabric Interest Group (active)
    - Data Publishing Interest Group (active; in collaboration with ICSU World Data System)
    - Data Type Registries Working Group (finished phase 1, starting phase 2)
    - Metadata Interest Group (active)
    - PID Information Types Working Group (finished)
    - Persistent Identifiers Interest Group (active)
    - Research Data Collections Working Group (active)
  - *Taverna (workflow management system)* http://www.taverna.org.uk/
  - *Thomson-Reuters* http://innovation.thomsonreuters.com
  - *THOR (and its precursor ODIN) project* http://project-thor.eu/
  - *W3C (World Wide Web Consortium)* https://www.w3.org
  - *Wf4ever project* http://wf4ever.github.io/ro/
- Webinars organized by RDA, OpenAIRE, THOR.
- Proceedings from some recent conferences (IEEE etc.),
- Articles in scientific literature (see bibliography)
- Discussions with colleagues and experts, from ENVRIplus partners and other organisations

Two-to-five year analysis

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

Here, we list some of the issues and ideas that are being worked on now, and which we feel will continue to be of importance in the coming years:

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

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

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

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

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

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

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

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

There is also active participation, by individual ENVRIplus RIs, in projects such as EUDAT2020 or as use cases in RDA groups. However, the relatively short lifetimes, and limited number of members, of this type of project often has several negative consequences. Firstly, there may not be enough diversity within the use cases to encourage the development of broad solutions that cover the needs and requirements of a wider range of communities. Secondly, the knowledge and experience gained through such work often ends up benefiting only a small number of RIs – if there is any long-lasting application at all!

- ENVRIplus could therefore make a difference by setting up a platform for informing practitioners about on-going initiatives (especially those that involve ENVRIplus members, but not as part of ENVRIplus itself), collection of RI use cases for passing on to the technology developers, and finally promoting the dissemination, implementation and uptake of effective examples.

Details underpinning above analysis

In this section, we present more background for the 7 topics (A-G) listed above. For each topic, some specific examples of relevant technologies are listed, together with a brief narrative discussion and suggestions for further reading – either links to the bibliography or to organisations whose web site addresses are listed under 4.2.2.

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

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

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

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

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

Sources: [Tilmes 2010], [Duerr 2011] (see example in the article supplement!)] + on-going work in RDA Metadata Interest Group, RDA Research Data Provenance Interest Group and EUDAT2020 (Work Package 8).

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

Currently, a majority of the ENVRIplus RIs – and their intended user communities – haven’t yet started to implement the outlined practices in a consistent manner. As a consequence, the reproducibility of research based on data from these RIs could be called into question. What is needed to change this situation, are good examples and demonstrators that can be easily adopted by the RIs (without much investment in time and software). Such best practices need to be developed in cooperation across the Work Packages of Theme 2.

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

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

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

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

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

- Systems for cataloguing and handling more complex collections, both of data sets and metadata (c.f. “research objects”).

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

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

In the simplest form, collections can consist of lists of individual data objects that belong together, such as 365 daily observations from a given year. Similarly, it may be desirable to combine data and associated metadata into packages, or to create even more complex “research objects” that may also contain annotations, related articles and reports, etc. Collections can be defined by the original data producers, but may also be collated by the users of the data – and may thus contain information from a large variety of sources and types. 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.

There is also a need for sustainable registries for data type definitions that can be applied to “tag” content in a way that is useful and accessible both to humans and for machine-actionable workflows. However, the use of data types varies greatly between different user communities, making it a difficult task to coordinate both the registration of definitions as well as a sustainable operation of the required registries, especially if these are set up and operated by RIs. Here more work is needed in collaboration with a number of RIs each with differing data-set structures and catalogue organisations, in order to provide clear recipes for data typing.

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

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

Sources: [Duerr 2011], [Socha 2012], [Klump 2015] + work by the RDA PID Information Types WG (concluded) and the RDA PID Interest Group (starting now).

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

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

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

Sources: RDA PID Information Types WG (final), new RDA Data Collections WG + presentations from the ePIC & DataCite PID workshop in Paris, 2015[1].

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

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

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

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

Sources: ORCID and DataCite, THOR web site and webinar series.

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

Connected with this is a growing trend to implement tighter information exchange (primarily links to content) between publishers, data repositories and data producers. There are several on-going initiatives looking into how to optimise and automate this, including the THOR project (operated by CERN), which involves amongst others OpenAIRE, ORCID, DataCite and Pangea. It is expected that the outcomes of these efforts will set the norm.
However, to be fully inclusive and consistent (from a data curation and cataloguing point of view), this practice should be extended to all relevant “personnel categories” involved in the research data life cycle, including technicians collecting data, data processing staff, curators, etc. – not just principal investigators and researchers, This would allow both a complete record of activities for individuals (suitable for inclusion in a CV), but conversely can also be seen as an important source of provenance information for linked data sets.

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

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

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

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

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

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

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

Sources: [Socha 2012], ePIC and DataCite PID workshop (Paris, 2015)[68], Make Data Count project, CrossRef, OpenAIRE, THOR.

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

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

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

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

Sources: RDA Data Type Registries Working Group, EUDAT

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

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

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

---

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

Sketch of a longer-term horizon

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

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

- A move towards automation of those aspects of the research data lifecycle that will involve basic tasks like assigning identifiers and citing or referring to all kinds of resources – including data and metadata objects, software, workflows, etc.

- Evolution towards more complex “collections” of research resources, like Research Objects, that will necessitate a more flexible approaches towards both strategies for identification and detailed, unambiguous citation or referencing parts of such objects.
- Much more tightly integrated systems for metadata, provenance, identification and citation will evolve (pushed by data producers, publishers and data centres), offering rapid and trusted feedback on data usage and impact.

Relationships with requirements and use cases

Requirements

There are strong connections between the RI requirements gathered for *identification and citation* with those related to other topics, including *cata loguing*, *curation*, *processing* and *provenance*. A majority of RIs are very concerned with how to best encourage and promote the use of their data products in their designated scientific communities and beyond, but at the same time, it is considered a high priority to implement mechanisms and safeguards that can ensure that the data producers (especially principal investigators and institutes in charge of data collecting and processing) receive proper credit and acknowledgments for their efforts. Here, it seems obvious that consistent allocation of persistent identifiers, and the promotion of standards for using these when citing data use in reports and publications, will go a long way to fulfil these needs. In addition, efforts to standardise the practices and recipes for identifying subsets of complex data collections, and subsequent extraction of micro-attribution information related to these subsets, would ensure a fair distribution of professional credit asked for by researchers and funding agencies alike.

Work packages

The overarching objective of the ENVRIplus Work Package 6 is to improve the efficiency of data identification and citation by providing recommendations and good practices for convenient, effective and interoperable identifier management and citation services. WP6 will therefore focus on implementing data tracing and citation functionalities in environmental RIs and develop tools for the RIs, if such are not otherwise available.

Use cases

Of the proposed ENVRIplus case studies, those of interest from an I&C perspective are mainly IC_01 “Dynamic data citation, identification & citation”, IC_06 “Identification/citation in conjunction with provenance” and IC_09 “Use of DOIs for tracing of data re-use”. (At the time of writing, these are under review or preparation, with some likelihood of a merger of the three.) The primary aim of IC_01 is to provide demonstrators of the RDA Data Citation Working Group’s recommendation [Rauber 2016] for a query-centric approach to how retrieval, and subsequent citation, of dynamic data sets should be supported by the use of database systems that track versions. This may be combined with support also for collections of data sets, which can be seen as a sub-category of dynamic datasets, thus addressing also the goals of IC_09. IC_06 is aimed at identifying good practices for using PIDs for recording provenance throughout the data object lifecycle, including workflows and processing.

Summary of analysis highlighting implications and issues

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

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

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

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

Further discussion of the data identification and citation technologies can be found in Section 4.2.5. This takes a longer term perspective and considers relations with strategic issues and other technology topics.

Bibliography and references to sources

[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.

[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. Available at http://link.springer.com/content/pdf/10.1007%2Fs12145-011-0083-6.pdf

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

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

[Klump 2015] J. Klump, R. Huber and M. Diepenbroek, ”DOI for geoscience data - how early practices shape present perceptions”. Earth Science Informatics Volume 9, Issue 1, pp 123-136, March 2016. Available via http://dx.doi.org/10.1007/s12145-015-0231-5.

[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/joint-declaration-data-citation-principles-final

[Moreau 2008] L. Moreau, P. Groth, S. Miles, J. Vazques-Salceda, J. Ibbotson, S. Jiang, S. Munroe, O. Rana, A. Schreiber, V. Tan and L. Varga, “The Provenance of Electronic Data”. Communications of the Association for Computing Machinery (ACM), volume 51, number 4, April 2008. Available at http://faculty.utpa.edu/fowler/csci6174/papers/Reilly_provenaceCACM.pdf.

[Parsons 2010] M.A. Parsons, R.E. Duerr and J.-B. Minster, ”Data citation and peer review”, EOS, Transactions of the American Geophysical Union vol 91, no 34, 24 August 2010, 297-304. Available at http://modb.oce.ulg.ac.be/wiki/upload/Alex/EOS_data_citation.pdf.

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

[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, Special Issue on Data Citation, 2016. (In press; pre-print available from author.)

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

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

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

[Tilmes 2010] C. Tilmes, Y. Yesha and M. Halem, “Tracking provenance of earth science data”. Earth Science Informatics 3:59-65, Volume 3, Issue 1, pp 59-65, June 2010. Available via http://dx.doi.org/10.1007/s12145-010-0046-3.

[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.

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