IC_1 Dynamic data identification & citation

1. Background

1.1 Short description

Identification of data (and associated metadata) throughout all stages of processing is really central in any RI. This can be ensured by allocating unique and persistent digital identifiers (PIDs) to data objects throughout the data processing life cycle. The PIDs allow unambiguous references be made to data during curation, cataloguing and support provenance tracking. They are also a necessary requirements for correct citation (and hence attribution) of the data by end users, as this is only possible when persistent identifiers exist and are applied in the attribution.
At the same time, research data is changing over time as new records are added, errors are corrected and obsolete records are deleted from a data set. Researchers rarely use an entire data set or stream data as it is, but rather create specific subsets tailored to their experiments. In order to keep such experiments reproducible and to share and cite the particular data used in a study, researchers need means of identifying the exact version of a subset as it was used during a specific execution of a workflow, even if the data source is continuously evolving.
In this implementation case we evaluate the requirements from the RI's gathered on the topics of Identification and Citation, and define the best candidates for technologies that will allow implementation of data citation for dynamic datasets and collections of datasets.

1.2 Contact

Background	Contact Person	Organization	Contact email
RI-Domain (Use Case Proposor, Agile Group Leader, WP6 leader)	Alex Vermeulen Maggie Hellström	ICOS	alex.vermeulen@nateko.lu.semargareta.hellstom@nateko.lu.se
RI-Domain	Christian Pichot	ANAEE	christian.pichot@paca.inra.fr
RI-Domain	Markus Fiebig	ACTRIS	Markus.Fiebig@nilu.no
RI-Domain	Barbara Magnana	LTER	Barbara.Magagna@umweltbundesamt.at
RI-Domain	Francois Andre	IAGOS	francois.andre@obs-mip.fr
RI-ICT	Markus Stocker	PANGAEA	mstocker@marum.de

1.3 Use case type

Implementation case

1.4 Scientific domain and communities

Scientific domain

Atmosphere | *Biosphere* | *Hydrosphere* | *Geosphere*

Community

Data Acquisition | *Data Curation* | *Data Publication* | *Data Service Provision* | *Data Usage*

Behavior

Data product generation/ Data Replication/Data Publication/Semantic Harmonisation/Data Discovery and Access/Data Citation

Roles

Data curator/Data publication repository/Service Provider

2. Detailed description

Objective and Impact

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

Challenges

The RDA recommendation on data citation requires that all metadata – and possibly also the data – is stored in the form of a versionable database. While this can be implemented relatively straightforwardly from scratch, a major reconstruction effort is required for existing metadata databases and /or flat file-based data storage approaches. Another major challenge is the requirement to guarantee that the database will be "future proof" and will also work 20 years from now supporting the same queries. Proper attribution also requires that citation services like Datacite support the harvesting of the contributor metadata in their citation indices. Other challenges are that this requires a mechanism to identify the uniqueness of queries and that all data is stored with stable sorting.
QUESTION: Do we want to add a few more sentences covering collections and the issues connected to providing proper citation (text snippets) for correctly attributing subsets of data extracted from larger datasets?

Detailed scenarios

Identify the database and query technologies that fulfill the requirements of the RDA recommendations for dynamic data. Select an implementation and test this on an existing (large) dataset for which sufficient metadata is available.
QUESTION: should we aim for two cases: 1) support for versioned flat-file-storage of data, and 2) support for true versionable data database systems?

Technical status and requirements

See above

Implementation plan and timetable

M1-3: Identification of technologies
M3-9: Setup of database and query system, generation of citation strings
M9-12: testing and refining

Expected output and evaluation of output

A show case implementation of the recommendations of the RDA working group on citation for dynamic datasets and collections of datasets.

External Links

1. IC_1 Notebook: {+}https://envriplus.manageprojects.com/projects/wp9-service-validation-and-deployment-1/notebooks/637+