AGU data citation community of practice - Credit for creators of data within collections using the concept of a reliquary

Justin Buck[1], Deb Agarwal[2], James Ayliffe[1], Chris Erdmann[3], Carole Goble[4], Ugis Sarkans[5], Daniel Noesgaard[6], Uwe Schindler[7], Shelley Stall[8], Martin Fenner[9], Martina Stockhause[10], and Paolo Manghi[11]

[1]National Oceanography Center
[2]Lawrence Berkeley National Laboratory
[3]AGU
[4]University of Manchester
[5]EMBL-European Bioinformatics Institute
[6]GBIF Secretariat
[7]MARUM - University of Bremen
[8]American Geophysical Union
[9]Front Matter
[10]German Climate Computing Centre (DKRZ)
[11]Consiglio Nazionale delle Ricerche (CNR)

November 22, 2022

Abstract

A gap in community practice on data citation that emerged during the AGU fall meeting 2020 Data FAIR Town Hall, "Why Is Citing Data Still Hard?" with the goal of addressing the use case of citing a large number of datasets such that credit for individual datasets is assigned properly. The discussion included the concept of a "Data Collection" and the infrastructure and guidance still needed to fully implement the capability so it is easier for researchers to use and receive credit when their data are cited in this manner. Such collections of data may contain thousands to millions of elements with a citation needing to include subsets of elements potentially from multiple collections. Such citations will be crucial to enable reproducible research and credit to data and digital object creators. To address this gap, the data citation community of practice formed including members from data centres, research journals, informatics research communities, and data citation infrastructure. The community has the goal of recommending an approach that is realistic for researchers to use and for each stakeholder to implement that leverages existing infrastructure. To achieve data citation of these subsets of large data collections the concept of a "reliquary" is introduced. In this context the reliquary is a container of persistent identifiers (PIDs) or references defining the objects used in a research study. This can include any number of elements. The reliquary can then be cited as a single entity in academic publications. The reliquary concept will enable data citation use cases such as the citation of elements within a data collection that are formed from numerous underlying datasets that have their own PIDs, unambiguous citation of data used in IPCC Assessment Reports, and citing the subsets of collections of research data that contain millions of elements. The discussions over the course of 2021 have developed a theoretical concept, at the time of writing formal use cases and initial applications are being defined. The recommendation developed by this effort will be available for review and comment by communities such as ESIP and RDA. All are welcome.

AGU fall meeting - 17th December 2021 - IN54A - Best Practices and Realities of Research Data Repositories I Oral

Data citation community of practice - Credit for creators of data within collections using the concept of a reliquary

Justin James Henry Buck (National Oceanography Center), Deb Agarwal (Lawrence Berkeley National Laboratory), James Ayliffe (National Oceanography Center), Chris Erdmann (AGU), Carole Goble (University of Manchester), Ugis Sarkans (EMBL-European Bioinformatics Institute), Daniel Noesgaard (GBIF Secretariat), Uwe Schindler (MARUM - University of Bremen), Shelley Stall (American Geophysical Union), Martin Fenner (Front Matter), Martina Stockhause (DKRZ German Climate Computing Centre), Paolo Manghi (Consiglio Nazionale delle Ricerche (CNR))

(Plus many more contributors to the workshops)

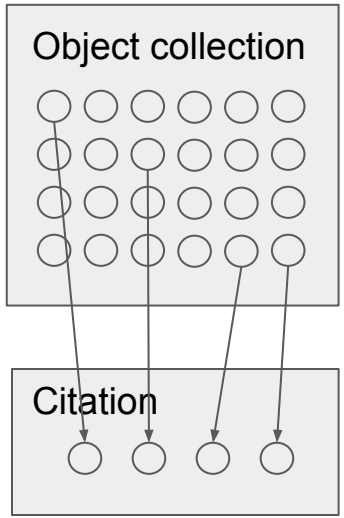
The challenge

- Enable citation of a large numbers of papers, software, and datasets (research objects) in a paper by providing a means to collapse them into a small number of references.

- Allow a project or data system collecting large numbers of research objects to enable citation of the group and constituents within it

- Empower an individual to create a group of research objects that might span repositories

- Provide an ability to cite any subset of research objects belonging to one or multiple groups



The data citation community of practice

Academic publishing

- AGU
- JATS4R
- Citation styles
- Outreach/guidance/training

Infrastructure

- DataCite
- CrossRef
- Scholix / OpenAire
- RO-Crate
- Zenodo
- Schema.org
- Web of Science

Community use cases

- RO-Crate
- BioStudies
- Global Biodiversity Information Facility (GBIF)
- PANGAEA
- Intergovernmental panel on climate change (IPCC)
- British Oceanographic Data Centre

Solution - primary goal of unambiguous citation

Data collection

Data collection

Data collection

Reliquary

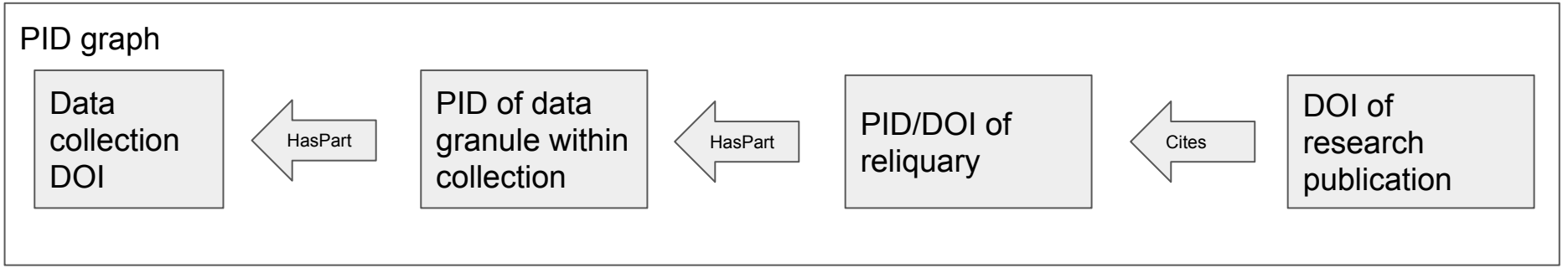
PID,ID_NAME/ID_DESC,UR(L/I/N),COMMENT
,RELIQUARY_CREATOR,R_CREATOR_TYPE
PID,ID_NAME/ID_DESC,UR(L/I/N),COMMENT
,RELIQUARY_CREATOR,R_CREATOR_TYPE
PID,ID_NAME/ID_DESC,UR(L/I/N),COMMENT
,RELIQUARY_CREATOR,R_CREATOR_TYPE
… one row for each entity included in citation …

(NOTE: example reliquaries currently being drafted for each use case)

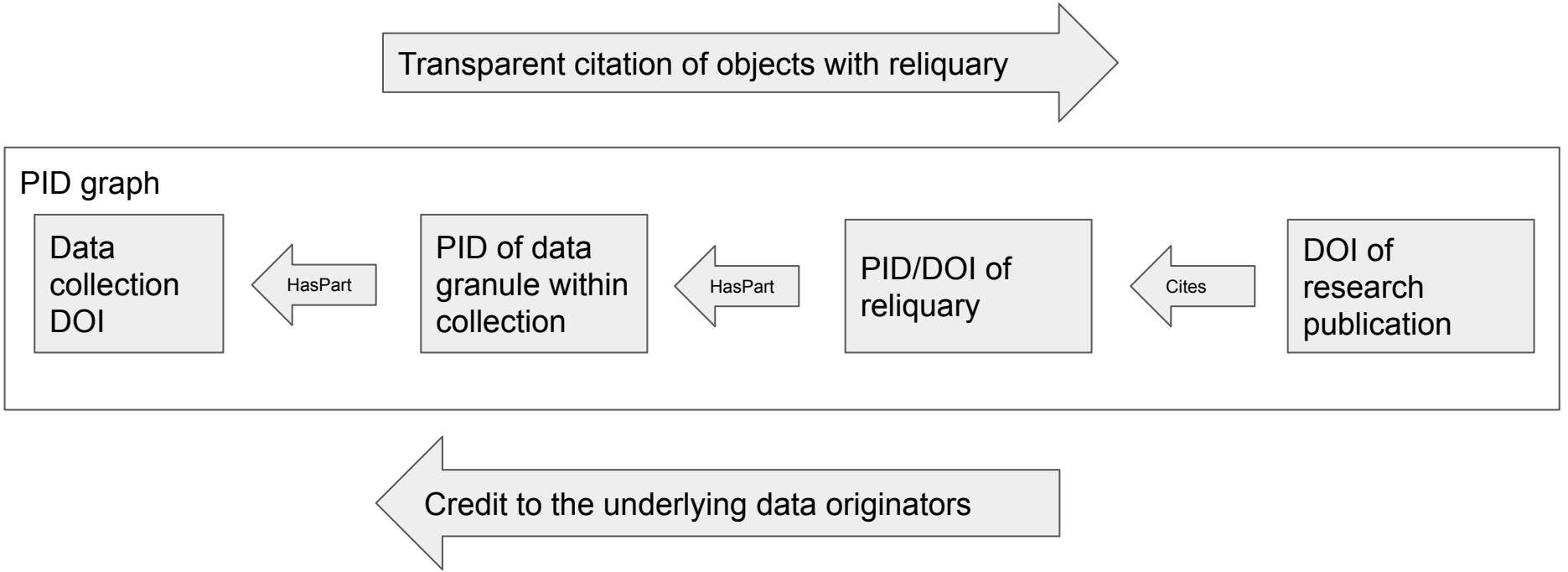
Each PID points to a landing Page

- Includes link to data
- Link may be brittle if data version is updated (pointer to new version?)
  - Reproducibility is important, current focus is transparency though

Dual role of reliquary



Dual role of reliquary



Next steps and how to engage with the community

Reliquary is the working title for the approach.

New research data alliance **Earth, Space, and Environmental Science Complex Citations** working group:

- https://www.rd-alliance.org/groups/earth-space-and-environmental-science-complex-citations-working-group

More information on the data citation community of practice site

- https://agu-data.github.io/DataCitationCoP/

If you have a use case or are keen to contribute to the working group we encourage you to get in touch.

Thank you

National Oceanography Centre
British Oceanographic Data
Centre BODC

Abstract here for notes but will not be in the final version

A gap in community practice on data citation that emerged during the AGU fall meeting 2020 Data FAIR Town Hall, “Why Is Citing Data Still Hard?” with the goal of addressing the use case of citing a large number of datasets such that credit for individual datasets is assigned properly. The discussion included the concept of a “Data Collection” and the infrastructure and guidance still needed to fully implement the capability so it is easier for researchers to use and receive credit when their data are cited in this manner. Such collections of data may contain thousands to millions of elements with a citation needing to include subsets of elements potentially from multiple collections. Such citations will be crucial to enable reproducible research and credit to data and digital object creators.

To address this gap, the data citation community of practice formed including members from data centres, research journals, informatics research communities, and data citation infrastructure. The community has the goal of recommending an approach that is realistic for researchers to use and for each stakeholder to implement that leverages existing infrastructure.

To achieve data citation of these subsets of large data collections the concept of a “reliquary” is introduced. In this context the reliquary is a container of persistent identifiers (PIDs) or references defining the objects used in a research study. This can include any number of elements. The reliquary can then be cited as a single entity in academic publications.

The reliquary concept will enable data citation use cases such as the citation of elements within a data collection that are formed from numerous underlying datasets that have their own PIDs, unambiguous citation of data used in IPCC Assessment Reports, and citing the subsets of collections of research data that contain millions of elements.

The discussions over the course of 2021 have developed a theoretical concept, at the time of writing formal use cases and initial applications are being defined. The recommendation developed by this effort will be available for review and comment by communities such as ESIP and RDA. All are welcome.