## Data Citation Community of Practice Workshop - 29 Oct 2021

Use Case: **Citing a large number of datasets such that credit for individual datasets works properly (second workshop)**

### Table of Contents (Bookmarks)

Information about Topic: <= Pre-reading and Orientation

[Notes from 8 April Initial Session](https://docs.google.com/document/d/1OGqqb6cOIZJiIEJ_zGzsMW5oldh4ioBdTl7JkcNkN7M/edit?usp=sharing) (link to document)

[Notes from 8 June Workshop](https://docs.google.com/document/d/18PpbBDp-7hWD4vyhLAjzVDnZt4xl_4lz4-NoYc3cV5Q/edit?usp=sharing) (link to document)

[Session Information](#eohv0rb54my) (Third)

[Abstract](#owuyw1t7lwpe)

[References](#_42lnvgftro9f)

[Examples](#_5s7yzjxk6v1h)

Third Workshop Details, 29 Oct: <= For Today’s Meeting navigate to this section

[**Agenda**](#s8ebr7m8kjzn)

[Code of Conduct](#gzgo92ne9gv2)

[Attendees](#vgf3113m33pz) <= Add your name here

[Collaborative Notes](#_nji0v2t7dlsz)

### Session Information

**Title:** Data Citation Community of Practice Workshopfor Data Citation

**When:** Oct 29, 2021 11:00 AM Eastern Time (15:00 UTC)

The data citation community of practice in the Earth, space, and environmental sciences (and related sciences) builds on discussions at the 2020 AGU Fall Meeting Data FAIR Town Hall, “Why Is Citing Data Still Hard?” as well as the initial workshop held on [8 April 2021](https://docs.google.com/document/d/1OGqqb6cOIZJiIEJ_zGzsMW5oldh4ioBdTl7JkcNkN7M/edit?usp=sharing) and follow on meeting [8 June 2021](https://docs.google.com/document/d/18PpbBDp-7hWD4vyhLAjzVDnZt4xl_4lz4-NoYc3cV5Q/edit?usp=sharing).

In this third workshop we share **three use cases on possible approaches to a “reliquary”** – a large collection of digital objects used to assist in citation and credit. These use cases come from work by the small team and now ready to share more broadly.

**Co-organizers:**

* Caroline Coward, NASA JPL - Leading the development of the Community of Practice
* Deb Agarwal, Lawrence Berkeley National Laboratory - Leading the scoping of the first use case and will be the use case “owner” for this working session.
* Shelley Stall - AGU Sr. Director for Data Leadership helping to organize and support the effort.
* Chris Erdmann - AGU Assistant Director for Data Stewardship helping to organize and support the effort.

[Data Citation Community of Practice Website](https://agu-data.github.io/DataCitationCoP/)

In order to give the recommendations generated from these workshops a place for review and support, we are **standing up an RDA working group** that will collaborate with other interest groups and working groups already established in RDA with interest in DOIs, DOI Collections, repositories creating DOI collections, journals supporting citing DOI Collections, and Credit for the elements of a DOI Collection. You are welcome to participate in these workshops and coordinate the development and review of the recommendations.

In addition to the RDA review process for recommendations, we will also ask members of **Earth Science Information Partners (ESIP)** for review and comment. Especially: COPDESS Cluster, Data Stewardship Committee, Discovery, Research Object Citation, and Sustainable Data Management.

### Abstract (established at the initial workshop):

A method is needed for handling citation of large numbers of objects, particularly datasets and software, in a scholarly work. The method should allow the citations of the objects to be counted as primary citations (as if they were cited in the reference section) in the scholarly work. Ideally, the solution also enables aggregators of large numbers of datasets to publish ‘*object collections*’ that can be cited together in the reference section of a scholarly work without losing the primary citation credit and tracking capability for the underlying object contributions. This need arose from a use case where projects were collecting large numbers (e.g. 70-370) of datasets each with their own authors and DOI but used together often in analyses (Agarwal 2021). In this use case, the dataset authors are not interested in being collected into a single larger dataset with a single identifier and the data usage rules request an individual citation for each underlying dataset.

### Problem:

1. Authors need to cite papers, software, and data in their papers. They are not always able to fit all the required citations in the reference section and thus end up citing many of them in the supplementary material. These references in supplementary materials are not currently counted as primary citations. They also do not print with the paper.
2. A project or data system that is collecting large numbers of datasets that might be used together needs a way to enable citation of the collection of datasets without violating the data usage policy requiring primary citation credit to the individual datasets used.
3. The references for the collection of datasets need to be interpretable even if the original repository no longer exists.

For the purposes of this document, we will generalize to talking about the solution as a need for an object collection where the collection can contain datasets, software, papers, etc and we will refer to the object collection as a ***reliquary***. A reliquary is defined as an object collection where precious objects (relics) are held for posterity. NOTE: We realize that reliquary may not be the right name for this container particularly given its religious connotation but, it has the advantage of not colliding with any already used term in this space.

### Ideal properties in a reliquary:

1. Typically organized around a theme (e.g. data and software used in an analysis, a large composite dataset describing a geographic region, a collection of a type of measurement, etc.)
2. Has a unique identifier and well defined citation components.
3. Each object in the reliquary has a unique identifier within the reliquary (ideally a globally unique identifier).
4. Each object in the reliquary has descriptive metadata providing, at a minimum, information needed to cite the object individually if needed.
5. The type and role of objects in the reliquary can be designated from well defined vocabularies.
6. The reliquary appears in the reference section of a paper and each citation of the reliquary should indicate what subset of the reliquary is being cited if it is not the whole reliquary.
7. Should be able to contain items from many sources and repositories.
8. Tracking of primary citation credit for objects in a reliquary is possible without additional help from the entity that created the reliquary. In addition, access to an object in the reliquary can be accomplished from the description of the object contained in the reliquary.
9. All objects in the reliquary have machine readable definitions, types, roles, and metadata. RO-Crate provides a potential prescription for the object definition)

### Related work:

[**RO-Crate**](https://www.researchobject.org/ro-crate/) - RO-Crate, is based around these principles: a) all metadata is Linked Data, using schema.org as much as possible; b) extensible for different domains; c) retain the core Research Object principles Identity, Aggregation, Annotation; d) inferred metadata rather than repetition; e) “just-enough” provenance; f) layered validation; g) archivable with BagIt and other packaging tools and compatible with digital preservation approaches; h) hooks to reuse existing domain formats; i) designed for easy programmatic generation and consumption. Similar to the approach of BioSchemas, rather than building new specifications from scratch, we aim to build best-practice guides and validatable profiles for building rich research data packages with existing standards, without requiring expert knowledge for developing producers and consumers. RO-Crate was developed as a way of gathering and pointing to datasets that are held in different repositories, transcending the scattered repository issue, and being able to exchange between repositories. RO-Crate used Linked Data for its metadata files.

(Sefton 2019, Goble 2021).

**Scalable dynamic data citations** - The approach recommended by the RDA Working Group relies on dynamic resolution of a data citation via a time-stamped query also known as dynamic data citation. It is based on time-stamped and versioned source data and time-stamped queries utilized for retrieving the desired dataset at the specific time in the appropriate version. Instead of providing static data exports or textual descriptions of data subsets, we support a dynamic, query centric view of data sets. The proposed solution enables precise identification of the very subset and version of data used, supporting reproducibility of processes, sharing and reuse of data. (Rauber 2015). Hunter and Hsu (2015) implemented this methodology to provide fine level credit to observers contributing to the eBird citizen science data set.

**Data collections** -

* Biostudies (https://www.ebi.ac.uk/biostudies/)
* NERC EDCs or any other environmental data centre (where DOIs are applied to data as they arrive which are then aggregated up in to harmonised data products/collections , also have the potential for high granularity PIDs within collections leading to a graph of PIDs that will also connect to other external PIDs)

**Data papers** -

Enhanced publications as a general means to define “dataset in context”

**Supplementary information** -

**Landing pages** - The internet accessible location that a PID resolves to that provides information about the object and access to the object.

**Data repositories** - Data repositories who serve collections of data from aggregate data collections containing multiple PIDs will need to serve data to users with RO crates to enable citation of data requested from the repository. Additionally, data repositories will need to serve the PID graphs of holdings (including PIDs, linkages/associations, and the graph/ontology) to PID/DOI citation service to enable metics/citations of individual PIDs to be determined (methods such as the RDA collections API may facilitate the sharing of the PID graph).

### References

(Agarwal 2021) Deborah A. Agarwal, Joan Damerow, Charuleka Varadharajan, Danielle S. Christianson, Gilberto Z. Pastorello, You-Wei Cheah, Lavanya Ramakrishnan, Balancing the needs of consumers and producers for scientific data collections, Ecological Informatics, Volume 62, 2021, 101251, ISSN 1574-9541, https://doi.org/10.1016/j.ecoinf.2021.101251.

<https://www.sciencedirect.com/science/article/pii/S157495412100042X?via%3Dihub>

(Sefton 2019) Sefton, Peter, Eoghan Ó. Carragáin, Carole Goble, and Stian Soiland-Reyes. "Introducing RO-Crate: research object data packaging." <http://conference.eresearch.edu.au/wp-content/uploads/2019/08/2019-eResearch_103_-Introducing-RO-Crate-research-object-data-packaging.pdf> (Presentation slides and blog post here: <http://ptsefton.com/2019/11/05/RO-Crate%20eResearch%20Australasia%202019/index.html>)

(Soiland-Reyes 2021) Stian Soiland-Reyes et al (2021): Packaging research artefacts with RO-Crate *In preparation* <https://stain.github.io/ro-crate-paper/>

(Goble 2021) Carole Goble, Stian Soiland-Reyes (2021): RO-Crate: A framework for packaging research products into FAIR Research Objects, <https://www.slideshare.net/carolegoble/rocrate-a-framework-for-packaging-research-products-into-fair-research-objects> RDA IG Data Fabric: FAIR Digital Objects 2021-02-25. [Video recording<https://www.youtube.com/watch?v=pz-MLdI7GLA>]

Hunter J., Hsu CH. (2015) Formal Acknowledgement of Citizen Scientists’ Contributions via Dynamic Data Citations. In: Allen R., Hunter J., Zeng M. (eds) Digital Libraries: Providing Quality Information. ICADL 2015. Lecture Notes in Computer Science, vol 9469. Springer, Cham. <https://doi.org/10.1007/978-3-319-27974-9_7>

(Rauber 2015) Andreas Rauber; Ari Asmi; Dieter van Uytvanck; Stefan Proell (2015): Data Citation of Evolving Data: Recommendations of the Working Group on Data Citation (WGDC). DOI: [10.15497/RDA00016](https://doi.org/10.15497/RDA00016).

<https://www.rd-alliance.org/group/data-citation-wg/outcomes/data-citation-recommendation.html>

Vannan, S., R. R. Downs, W. Meier, B. E. Wilson, and I. V. Gerasimov (2020), Data sets are foundational to research. Why don’t we cite them?, Eos, 101, <https://doi.org/10.1029/2020EO151665>. Published on 16 November 2020.

Bardi, A. and Manghi, P., 2014. Enhanced Publications: Data Models and Information Systems. LIBER Quarterly, 23(4), pp.240–273. DOI: <http://doi.org/10.18352/lq.8445>

Buneman,P., Christie,G., Davies, J.A. et al. Why data citation isn’t working, and what to do about it. Database (2020) Vol. 2020: article ID baaa022; doi:10.1093/database/baaa022

Dorta-González, P., González-Betancor, S.M. & Dorta-González, M.I. To what extent is researchers' data-sharing motivated by formal mechanisms of recognition and credit?. *Scientometrics* 126, 2209–2225 (2021). https://doi.org/10.1007/s11192-021-03869-3

### Examples:

From Martin Fenner: **Collection of Datasets that are registered in more than one repository**

The pattern I have seen - and that make sense to me - is to have a data repository register a collection of datasets as an intermediary if the number of datasets that otherwise needs to be cited directly by the publication exceeds a reasonable number (maybe 10-50, depending on journal policies and community norms). Not many data repositories allow the registration of collections with at least part of the content registered elsewhere. One example in the life sciences that focuses on this use case is Biostudies (<https://www.ebi.ac.uk/biostudies/>).

DataCite metadata would allow connecting all datasets to the publication via the collection as intermediary, we can do these 2-step connections in DataCite Commons.

From Juliane Schneider: Sage Bionetworks currently handles this by creating a ‘landing page’ with a DOI for a manuscript with links to the studies, the specific data from the studies, and access information (much of our data is human data, so requires a signed data use certificate for access). We also have records for experimental and computational tools, so links to those would also be included if relevant. (We’re currently working on creating a doi for a query that gathers all of the data from the various studies for a manuscript, possibly negating the need for the landing page but we’ll see how that pans out.)

Ex: <https://www.synapse.org/#!Synapse:syn23521805>

We then send the author a data accessibility statement that includes the landing page doi to be used in their manuscript:

<data, analysis output, tools (describe content)> are available via the AD Knowledge Portal ([https://adknowledgeportal.org](https://adknowledgeportal.synapse.org)). The AD Knowledge Portal is a platform for accessing data, analyses, and tools generated by the Accelerating Medicines Partnership (AMP-AD) Target Discovery Program and other National Institute on Aging (NIA)-supported programs to enable open-science practices and accelerate translational learning. The data, analyses and tools are shared early in the research cycle without a publication embargo on secondary use. Data is available for general research use according to the following requirements for data access and data attribution ([https://adknowledgeportal.org/DataAccess/Instructions](https://adknowledgeportal.synapse.org/DataAccess/Instructions)).

For access to content described in this manuscript see: <manuscript landing page DOI>"

From Kristin Ulmer: This is an example of a use case, not necessarily a solution (unfortunately). The [Next Generation Liquefaction](http://www.nextgenerationliquefaction.org/about/index.html) (NGL) project is creating an open, collaborative, dynamic [database](http://www.nextgenerationliquefaction.org/) of case history data. We cite hundreds of individual [references](http://nextgenerationliquefaction.org/citations.php) We rely heavily on others who collected the field data or made some significant interpretation of field data. We feel that it is right for those original sources of data to be cited by those who later use the NGL database (as a whole or a subset of it). But the list of citations becomes too long for most journal publishers (and citations in supplementary files do not get indexed, from what I understand). We want to identify best practices and provide guidance to the users of our database so that they will be more likely to cite the original data sources (in addition to the NGL database DOI). As an aside: the database can be accessed through Jupyter notebooks, and the table of citations (DOIs, text descriptions, text version of the citation) can be queried. Each citation is linked to one or more objects in the NGL database by primary/foreign keys (via SQL).

From Ian Bruno (CCDC): We have a collection of over 1 million repository (or database) records, each representing a single dataset typically deposited with (or sometimes abstracted by) us. Whilst these datasets can each be viewed and downloaded individually, true power comes from their aggregation, and the application of knowledge derived from that aggregation. This knowledge can be influential in designing new drugs and novel materials. Contributing to each "knowledge point" are individual data points taken from or derived from a subset of the original datasets. This subset could include hundreds or thousands of datasets. It would be great to be able to link back from knowledge that had impact elsewhere to the individual datasets that contributed to this, and recognise the contribution of those who published and shared these. This said, we don't necessarily have a community crying out for this so at this stage it is more of a thought experiment but, as a repository wanting to encourage sharing of data, it could be useful for demonstrating the impact an individual dataset can have when aggregated with others as part of a larger collection.

From Martina Stockhause - IPCC Data Repository (June 11, 2021)

So our use case are Figure Data reliquaries:

* They consist of model data, observations, reanalysis data (input datasets) and displayed trend lines of mean and other statistical values, which are created by the authors (final datasets). I think we should keep the input and final datasets in two reliquaries and don't mix them into a single one. At the moment we do dataset lists and citations of input datasets (if we are able to get them together until end of July 2021).
* Content of the input data reliquaries do not all have DOIs and some have one on a different granularity.
* CMIP6 model data special case: We have HandleIDs on individual datasets but for citation, DOIs are assigned on data of a model run performed by a modeling group. In terms of credit and data use analysis (data references in papers) that makes sense, as these datasets are a contribution to a model experiment. However, for the reliquary content these datasets have a HandleIDs and are part of a couple of DOIs. (Datasets are created by different modeling groups and in a figure a subset of the data contributed by several modeling centers is used, e.g. only temperature and precipitation out of model run data of 10 modeling centers - belonging to 10 DOIs).

Comment: As data licenses are often CC-BY it is required that the providers of the reliquary content get credit. We are having a long CMIP6 license discussion between IPCC and WGCM-CMIP.

### Agenda:

1. [8 min] Welcome and Introductions (Shelley Stall/Deb Agarwal)

**“Reliquary” Use Cases:**

1. British Oceanographic Data Centre, Justin Buck and James Ayliffe ([slides](https://docs.google.com/presentation/d/1hIzPEn8UhPKgjraRGJAdftnr94DIH6-pUQ956Z613Ks/edit?usp=sharing))
   1. [slides](https://docs.google.com/presentation/d/1zAMZcMOGbE3WtgpWEjUuloySeTG2W-W9fHQ4t2sTuIM/edit#slide=id.ge509643c53_0_246) (link to slides)
2. German Climate Computing Center/IPCC DCC, Martina Stockhause ([slides](https://docs.google.com/presentation/d/13d9ZTM5vZGkN__l4bsRaHSReZjb6G8st))
3. Ameriflux, Deb Agarwal ([slides](https://docs.google.com/presentation/d/1nBEuOLWFTNW7OqS1LZyfVuILJcRRuHUMom_beehRlzc/edit?usp=sharing))
4. [15 min] Discussion/Next Steps (Moderator: Deb Agarwal/Shelley Stall)
   1. **RDA Session**: 3 November, Breakout 3, 22:30 - 00:00 UTC [ESIP/RDA Earth. Space, and Environmental Sciences (ESES) IG](https://www.rd-alliance.org/groups/esiprda-earth-space-and-environmental-sciences-ig), [New Developments in ESES Infrastructure and Repositories](https://www.rd-alliance.org/rdas-18th-plenary-draft-programme)
   2. **AGU Session**: IN52B - 04, eLightening, **Abstract Title:** Enabling Citations of Large Numbers of Diverse Datasets, **Session Date and Time:** Friday, 17 December 2021; 09:45 - 11:00 CST
5. [5 min] Wrap Up (Deb Agarwal / Shelley Stall)

### Code of Conduct: <https://www.agu.org/Plan-for-a-Meeting/AGUMeetings/Meetings-Resources/Meetings-code-of-conduct>

### Attendees (29 Oct 2021):

| Name | Affiliation | Email |
| --- | --- | --- |
| Shelley Stall | American Geophysical Union | [sstall@agu.org](mailto:sstall@agu.org) |
| Deb Agarwal | Lawrence Berkeley National Laboratory | [daagarwal@lbl.gov](mailto:daagarwal@lbl.gov) |
| Chris Erdmann | American Geophysical Union | [cerdmann@agu.org](mailto:cerdmann@agu.org) |
| James Ayliffe | British Oceanographic Data Centre | [jamayl@bodc.ac.uk](mailto:jamayl@bodc.ac.uk) |
| Carole Goble | The University of Manchester UK  ELIXIR, RO-Crate, EOSC-Life | carole.goble@manchester.ac.uk |
| Martin Fenner | Front Matter | martin@front-matter.io |
| Martina Stockhause | German Climate Computing Center / IPCC DDC | stockhause@dkrz.de |
| Howard Ratner | CHORUS | hratner@chorusaccess.org |
| Maggie Hellström | ICOS Carbon Portal & Lund university (Sweden) | margareta.hellstrom@nateko.lu.se |
| George Porter | Caltech | george@library.caltech.edu |
| Justin Buck | NOC BODC | juck@bodc.ac.uk |
| Rachel Proudfoot | University of Leeds | r.e.proudfoot@leeds.ac.uk |
| Stian Soiland-Reyes | The University of Manchester UK  ELIXIR, RO-Crate, EOSC-Life | soiland-reyes@manchester.ac.uk |
| Lorraine Hwang | CIG, UC Davis | ljhwang@ucdavis.edu |
| Erin Antognoli | USDA National Agricultural Library | Erin.Antognoli@usda.gov |
| Elisha Wood-Charlson | Berkeley Lab, KBase/NMDC | elishawc@lbl.gov |
| Susan Borda | Univ of Michigan Library | sborda@umich.edu |
| Renée F. Brown | McMurdo Dry Valleys LTER | rfbrown@unm.edu |
| Danielle Christianson | Lawrence Berkeley National Laboratory | dschristianson@lbl.gov |
| Nancy Ritchey | NOAA NCEI | nancy.ritchey@noaa.gov |
| Caroline Coward | NASA Jet Propulsion Laboratory/Caltech | caroline.m.coward@jpl.nasa.gov |
| Myranda Uselton | NOAA NCEI | myranda.uselton@noaa.gov |
| Janete Saldanha Bach | GESIS – Leibniz Institute for the Social Sciences | janete.saldanhabach@gesis.org |
| Bruce E. Wilson | ORNL DAAC | wilsonbe@ornl.gov |
| Oliver Bandel | ZBMED (Germany) | Bandel@zbmed.de |
| Dudee Chiang | Jet Propulsion Laboratory Library | dudee.chiang@jpl.nasa.gov |
| Mark Parsons | University of Alabama in Huntsville | map0046@uah.edu |
| Sarah Siddiqui | University of Rochester | ssiddiqui@library.rochester.edu |
| Alex May | Oak Ridge National Laboratory | mayab@ornl.gov |
| Christine Laney | Battelle - National Ecological Observatory Network (NEON) | claney@battelleecology.org |
| Paige Morkner | National Energy Technology Laboratory | paige.morkner@netl.doe.gov |
| Liz Krznarich | DataCite | liz.krznarich@datacite.org |

### Collaborative Notes:

Reliquary - is a container for [relics](https://en.wikipedia.org/wiki/Relic) (<https://en.wikipedia.org/wiki/Reliquary>) Term being used for a collection of (links to) research objects (e.g., data, software, workflows, SOPs etc).

Collection of DOIs, but realize that there are additional identifiers/links to consider, other objects including software, RRIDs, IGSNs, lions, tigers, bears… oh my

**British Oceanographic Data Centre,** Justin Buck and James Ayliffe ([slides](https://docs.google.com/presentation/d/1hIzPEn8UhPKgjraRGJAdftnr94DIH6-pUQ956Z613Ks/edit?usp=sharing))

First talk:

What granularity to cite the data (instruments, curse data, etc)

Data accessions from scientists, reformat discrete measurements, DOI on every granule would be difficult (crazy), DOIs on collections from teams/scientists, point of submission, sort of like PIDs internally at granular level (accession internal term, receipt of data/date, where get and archive data in system), different PIDs parts of different collections, many to many relationships w/ PIDs in multiple collections, not linear

Query based on temporal, spatial level form systems

Want way to aggregate data, package, traceability and citability

Accession is used in the biosciences for a submission (the big bioscience databases have "accession numbers" as identifiers for their submissions)

Accession - “a new item added to an existing collection of books, paintings, or artefacts.

"the day-to-day work of cataloguing new accessions"” - from the library community (definition, Oxford Dictionary)

Maggie: Use of PIDs, DOIs, are they the same? DataCite DOIs are unsustainable for

granularity, looking at handles internally - DOIs are PIDs, also handles are PIDs - clarity on terminology - DOIs more metadata, vs more minimal with handles

DOIs are a subset of the PIDs used. They are all PIDs

Mark: Clarity on granular data, collection level

Maggie Hellström: unambiguous referencing at "granule level" seems very challenging! what happens if there are different versions involved - can potential reuse (and reproducibility) be guaranteed? Are end user communities aware of all the details involved, and how do you train them if needed?

Breakability is a feature (brittleness per versioning)

Transparency is a prerequisite

Rich metadata (e.g., DataCite DOIs) vs thin metadata (e.g., assoc with handles)

Internal identifiers/thin metadata may not support reproducibility (per Maggie)

Shelley: MVP has to have other identifiers beyond just DOIs

Reyna: We’re doing camera data similarly - one DOI per deployment, and then identifiers for each file.

Stian Soiland-Reyes:

In particular with intermediary PID with "thin" metadata it is important to propagate DOI-level authorship metadata onwards like in the RO-Crate. As it may be trickier to resolve that later (and expensive if there are thousands of them)

I think what Maggie says here with "intermediate fluctual local IDs" **will rather be the norm** than actual PIDs that will keep working globally. So this James told us is more of a "best case" scenario for complex collections. So we should keep either case of identifiers, even if they may not be easily resolvable. we can encourage UUID use as base case, so at least they are globally unique compared to "dataset 42"

From Maggie Hellström:

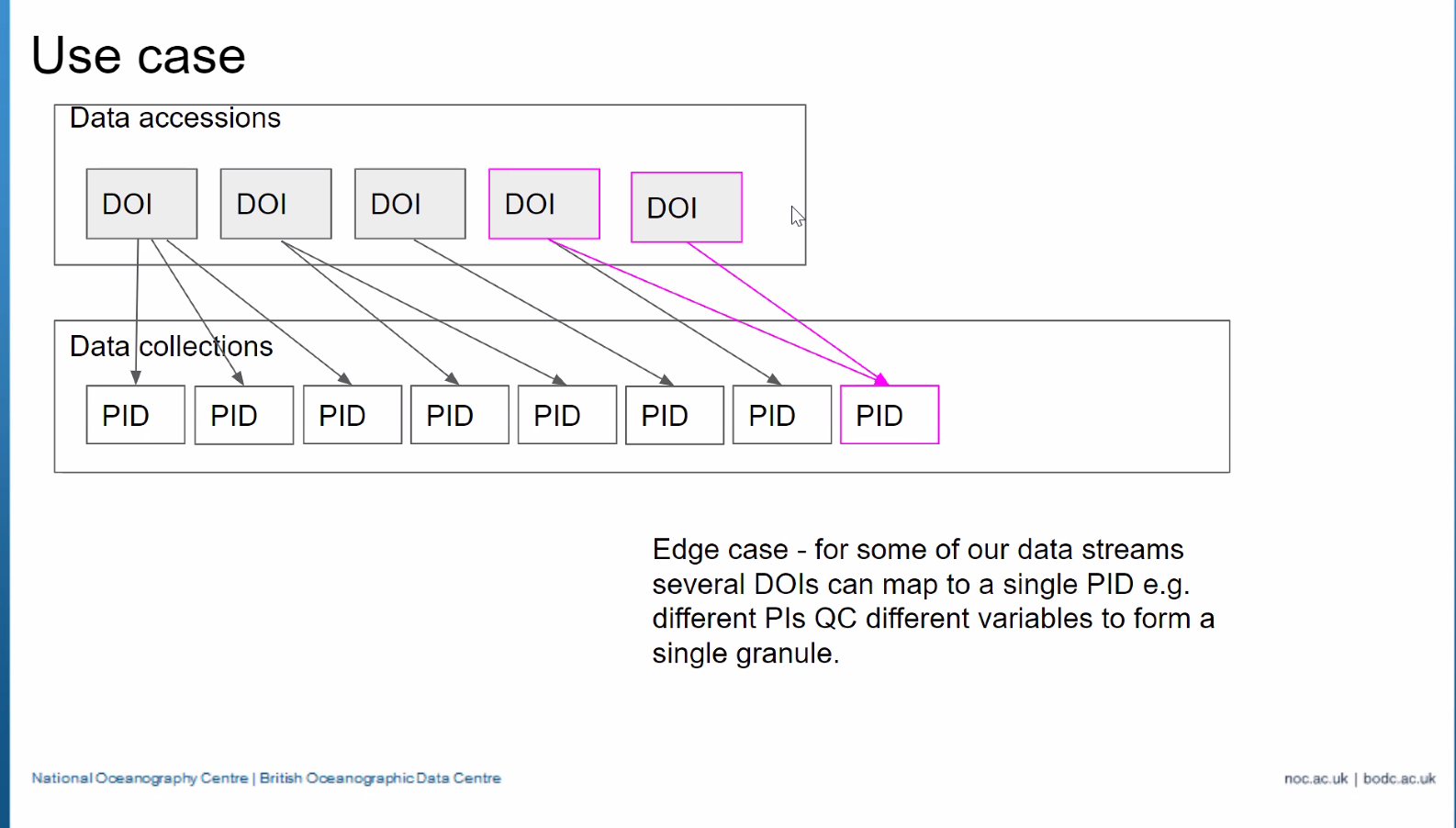
@Stian: no, you misunderstand me. Any "referenceable" DO should be given a GURPI that is indeed sustained and remains resolvable (although possibly to a tombstone) "forever".

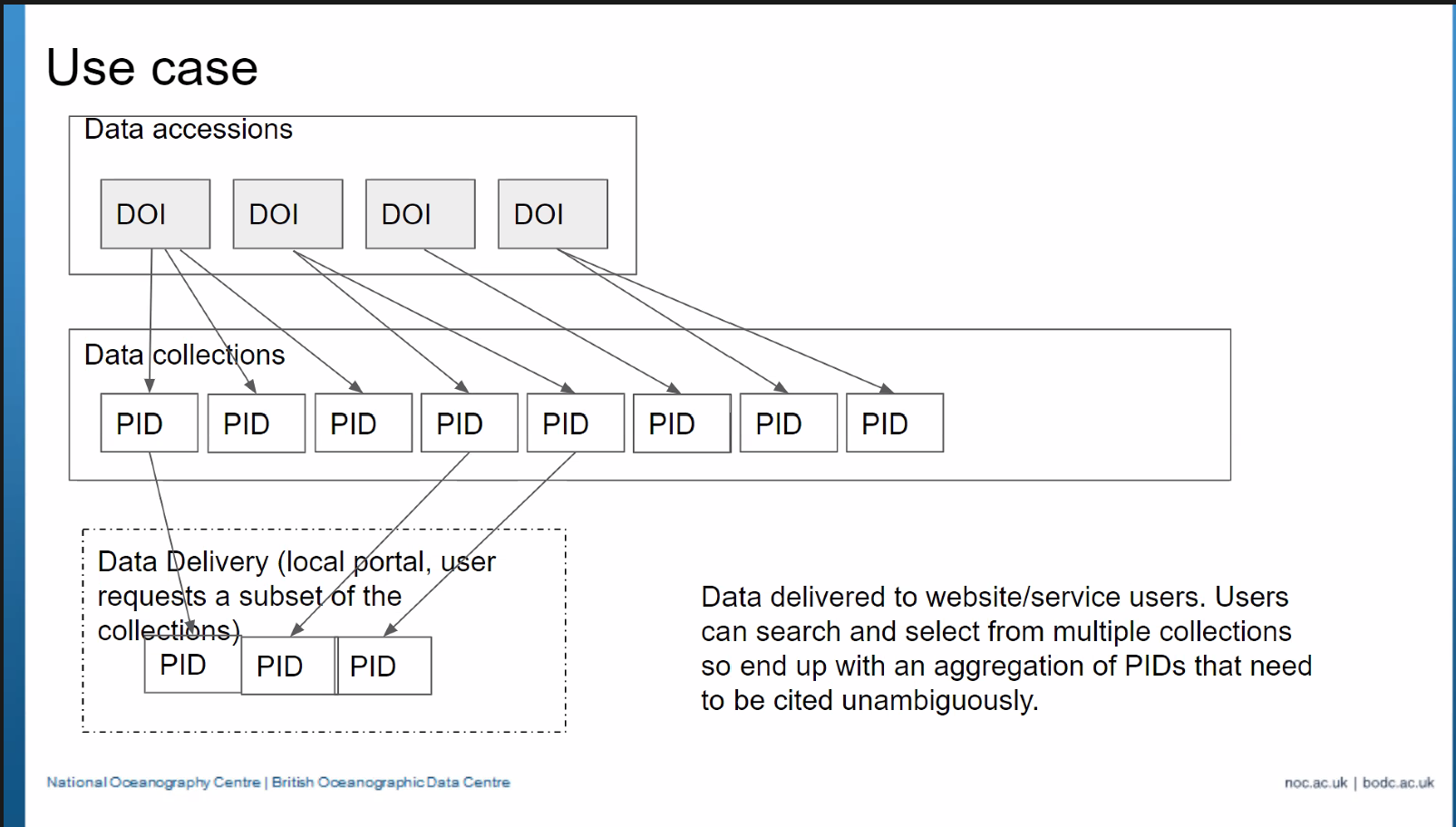
(Compare the Chinese (?) initiative to assign GURPIs to each individual packet of milk that is sold.)

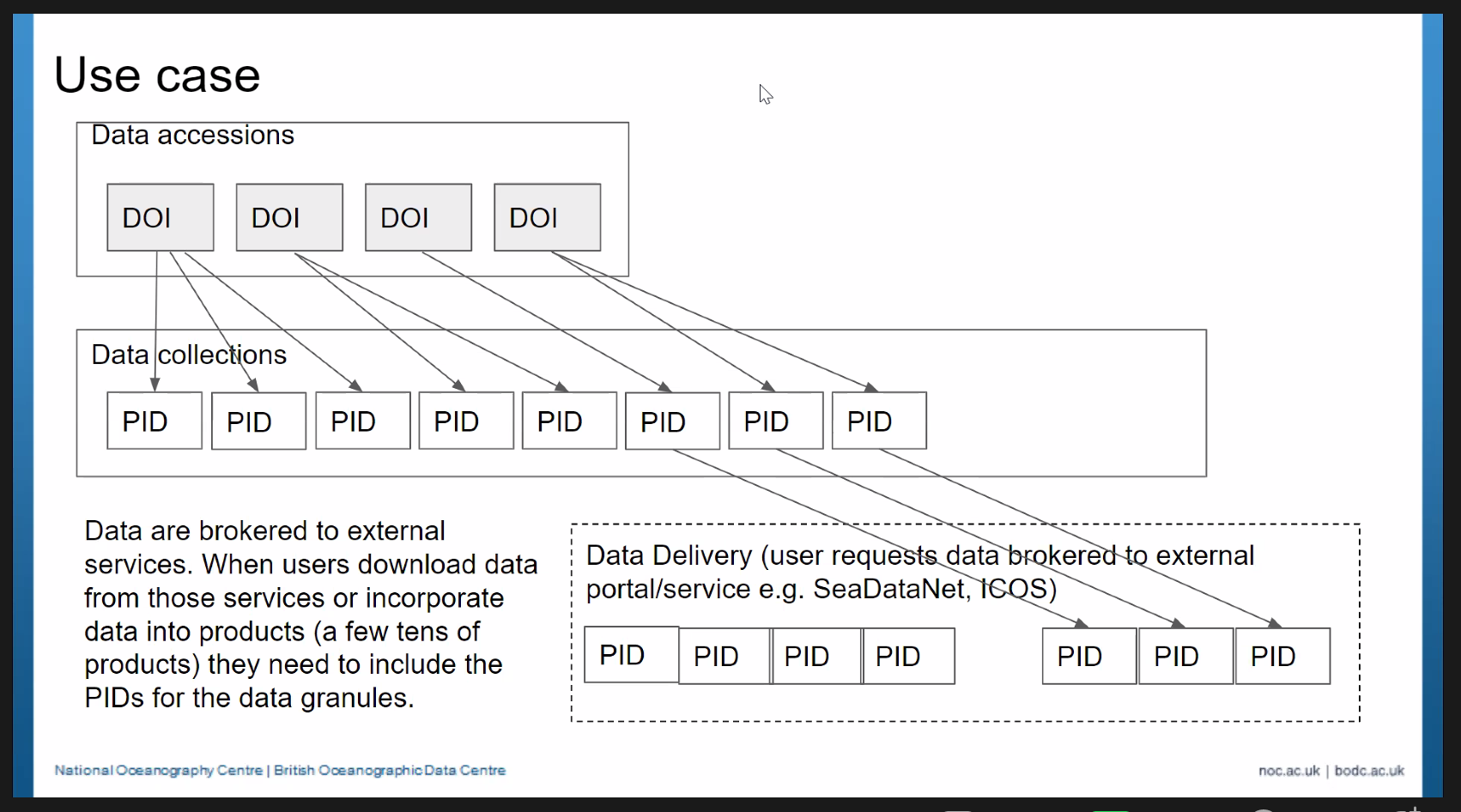
From Stian Soiland-Reyes :

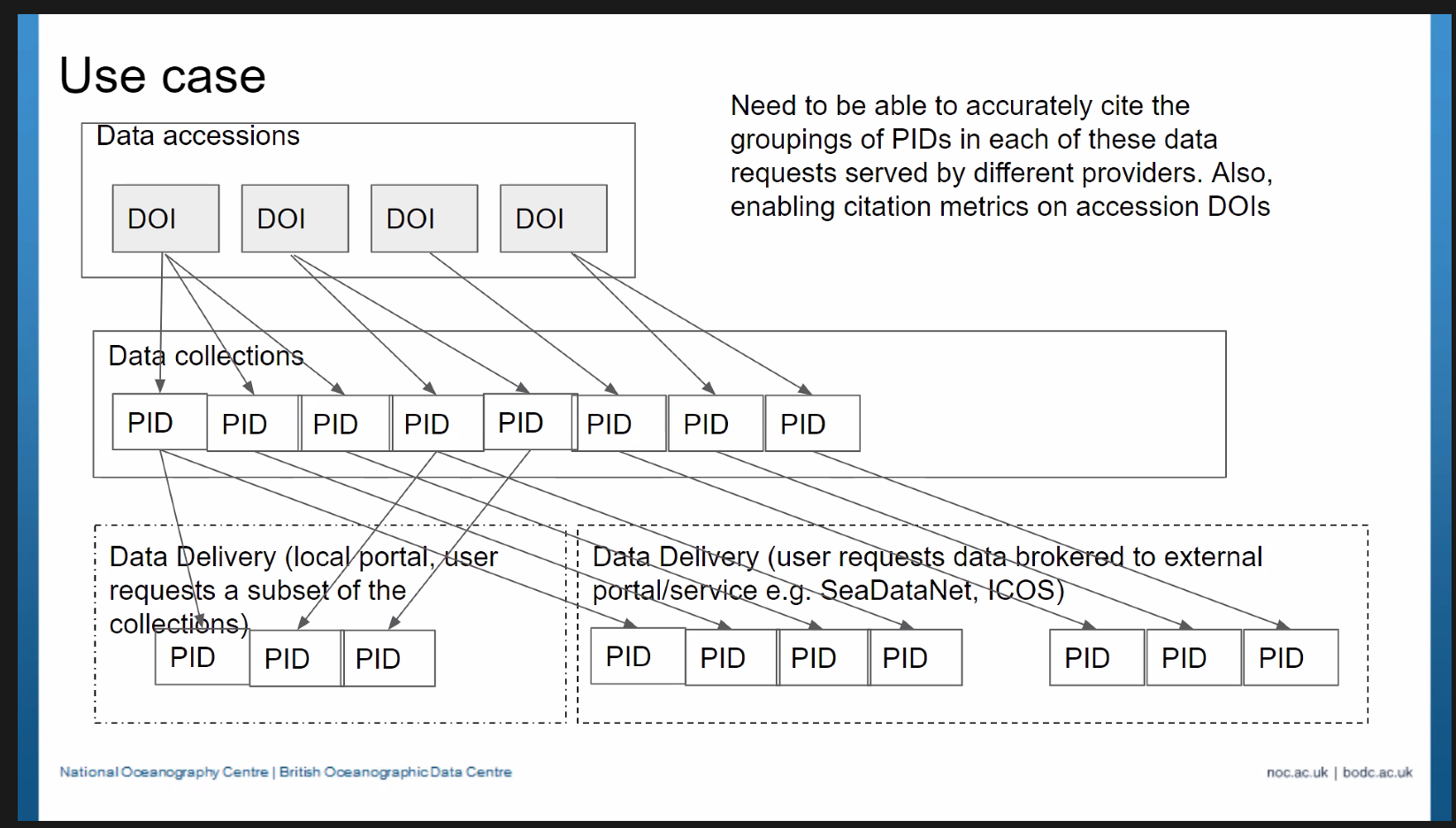
I think the reliquary may need to help as an intermediary for minting these on demand so that the people citing can fix/keep it even when abandoned by the original data provider ("New website design, let's break all URLs")

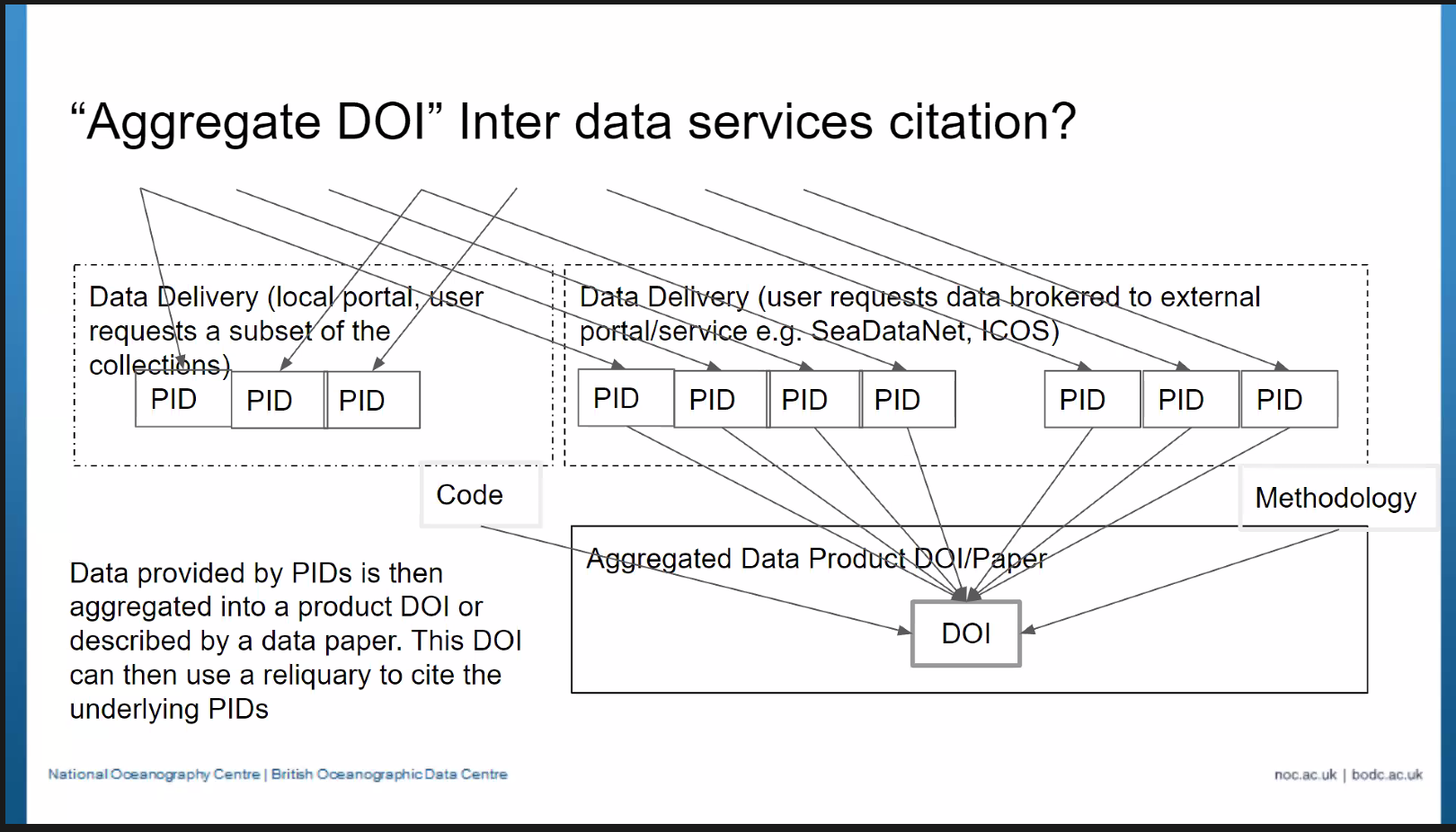
Examples: Datanet, EMO



****

****

****

****

**German Climate Computing Center/IPCC DCC,** Martina Stockhause ([slides](https://docs.google.com/presentation/d/13d9ZTM5vZGkN__l4bsRaHSReZjb6G8st))

Work connected to IPCC FAIR guidelines, nodes, working groups, coordination group

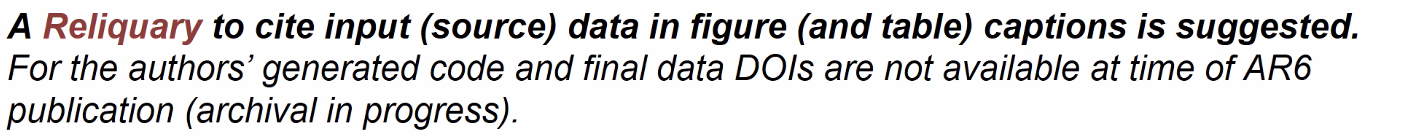
AR6 Report - Transparency, traceability of results, reproducibility of figures

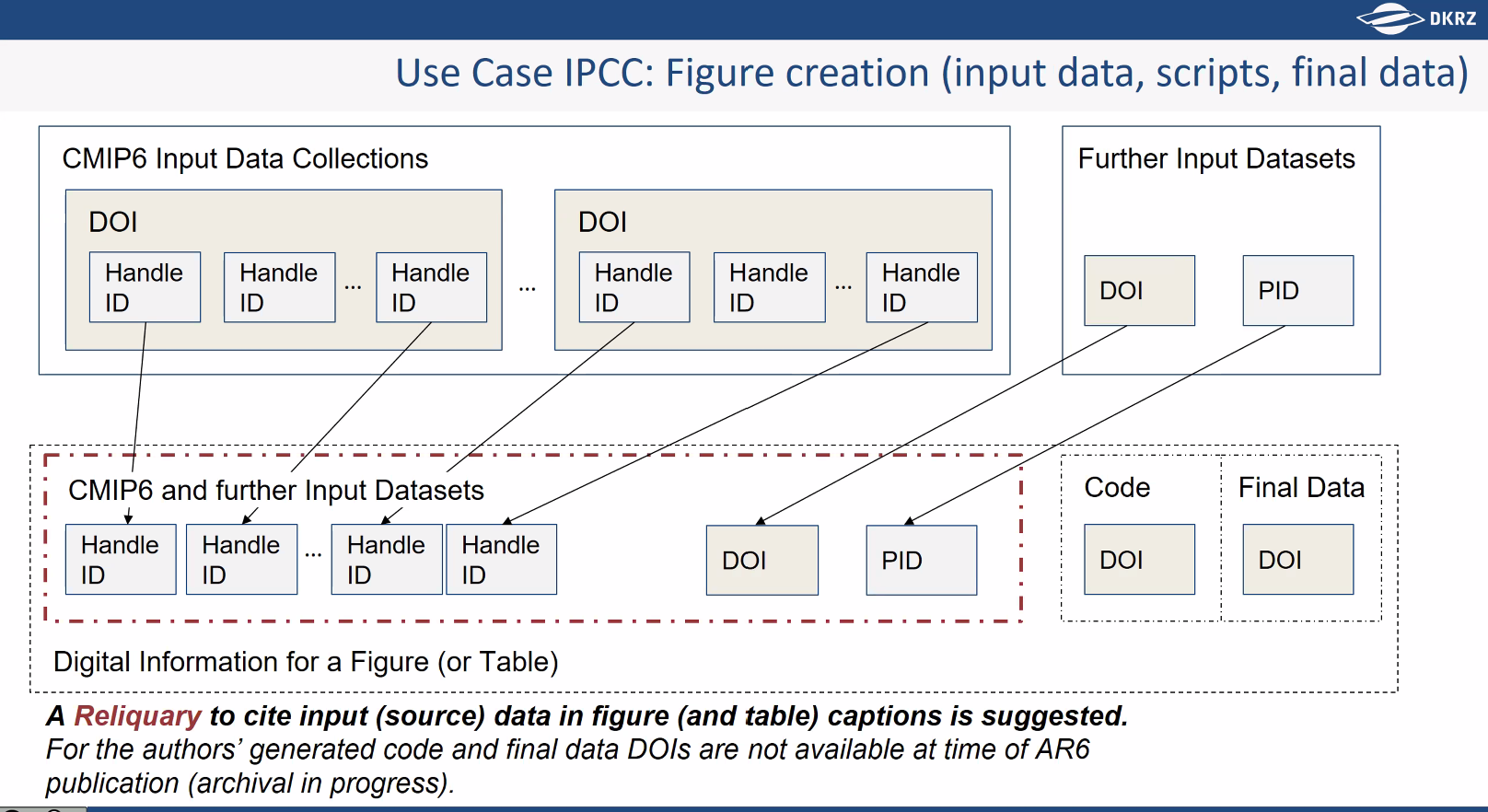
CMIP data input structure / data provider view

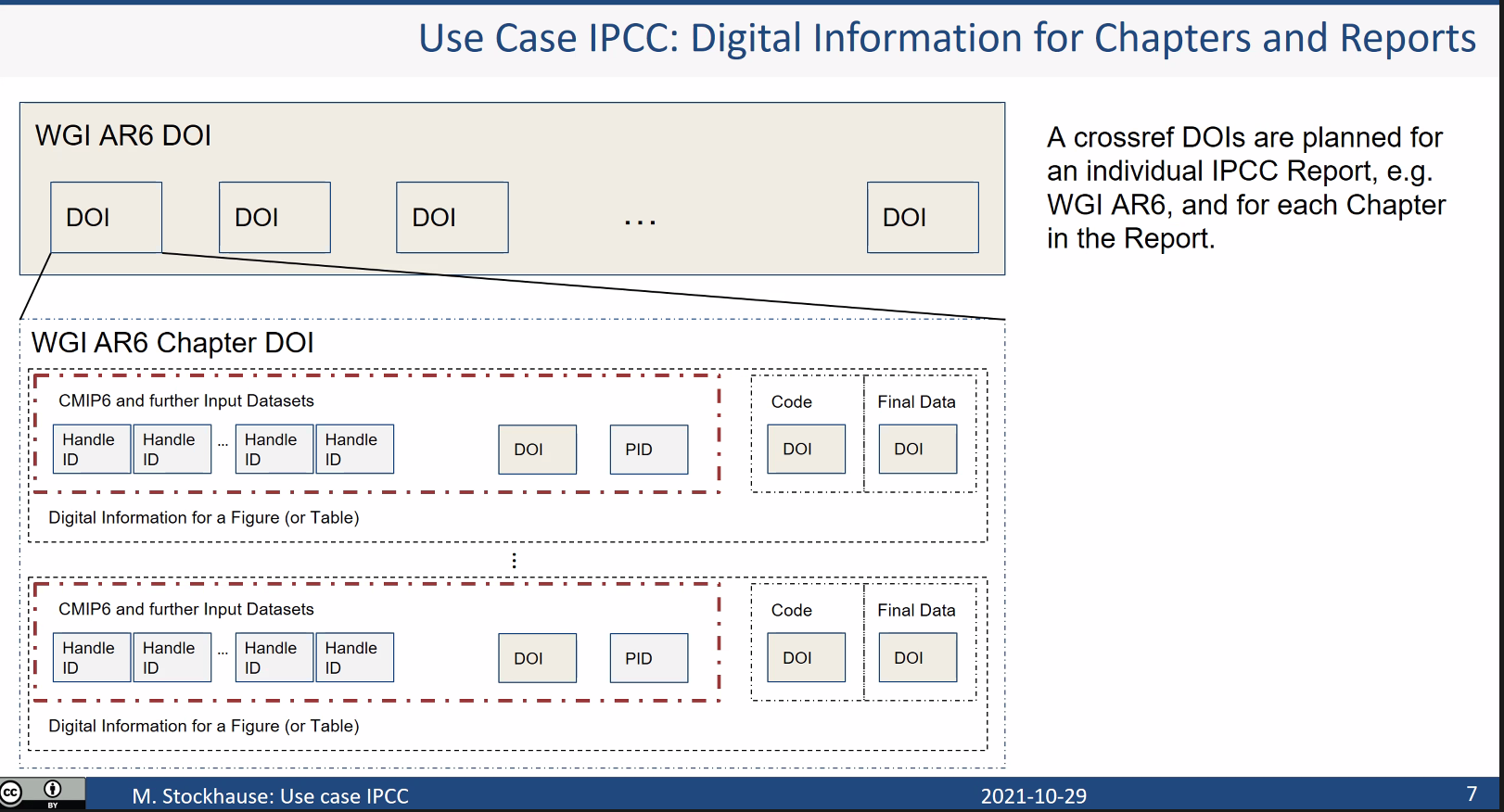
Model data - model runs, cite data together, time series climate parameters, sub granularity, time series can be one, many ids/files

Use data from different ensemble runs / select few handles out of DOIs, combine into study, to produce your own results, other input data will be added to study

Creation of figures, granularity, full chain, analyzer script, final data - reliquary on input datasets - this is created by authors





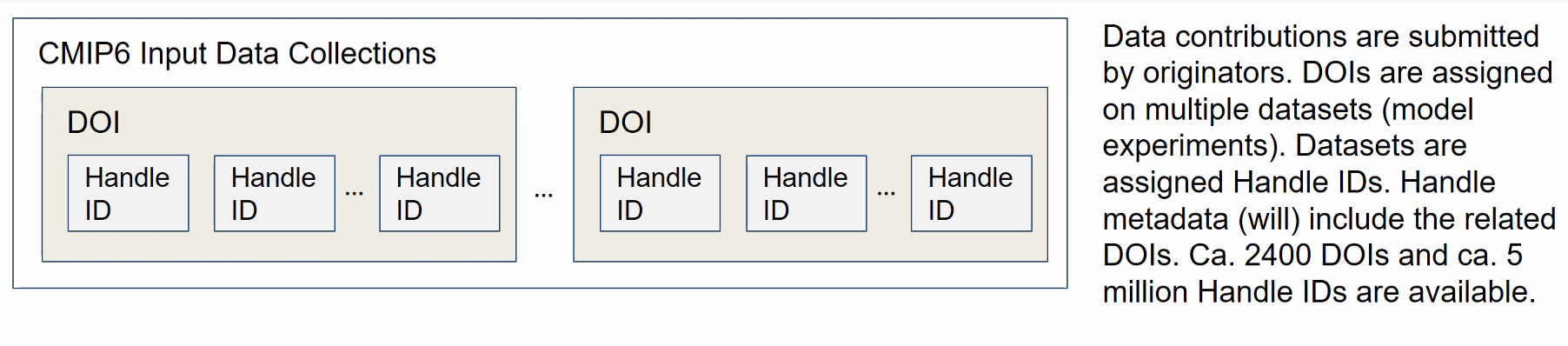


From Reyna Jenkyns to Everyone 04:36 PM

Looks like these handle IDs almost double as a query PID for the DOI that is pulled into the reliquary.

From Stian: Reyna, yes, query PIDs makes sense. Almost like OAI-ORE has a "Proxy" object for representing "this item as aggregated in this collection". Allows assigning alternative collection-specific titles for instance.

Concept of query PIDs vs concept presented by Martina



Stian: Unspecified vs specified selections/collections are both useful

Deb: Per Mark Parsons, original definition of reliquary has expanded based on these use cases. Different goals: 1) Proper citation (reliquary doesn’t have to be tied to pub) - pubs limited ability to properly cite multiple datasets 2) Some ability to give clarity of granularity/granules that were used -> Secondary goal has become a core goal, needs to be part of the accountability per Martina’s use case - How can we rethink the reliquary so that it is not too complex. Citation aspect, provenance part (maybe separate?). Originally - enable primary citation of elements in paper

Caroline: Hone in on definition of reliquary, credit, maintenance (q: is there mutability?) -> Create a smaller group/task force to tackle definition

Justin: Discussions are helpful for framing at AGU where DCCOP will be presented/discussed

From Martina Stockhause to Everyone 05:07 PM

What the concept of the reliquary could do for our use case is to **harmonize the reproducibility (based on data granules) together with the credit idea** on our large data collections.

**Ameriflux, Deb Agarwal (**[**slides**](https://docs.google.com/presentation/d/1nBEuOLWFTNW7OqS1LZyfVuILJcRRuHUMom_beehRlzc/edit?usp=sharing)**)**

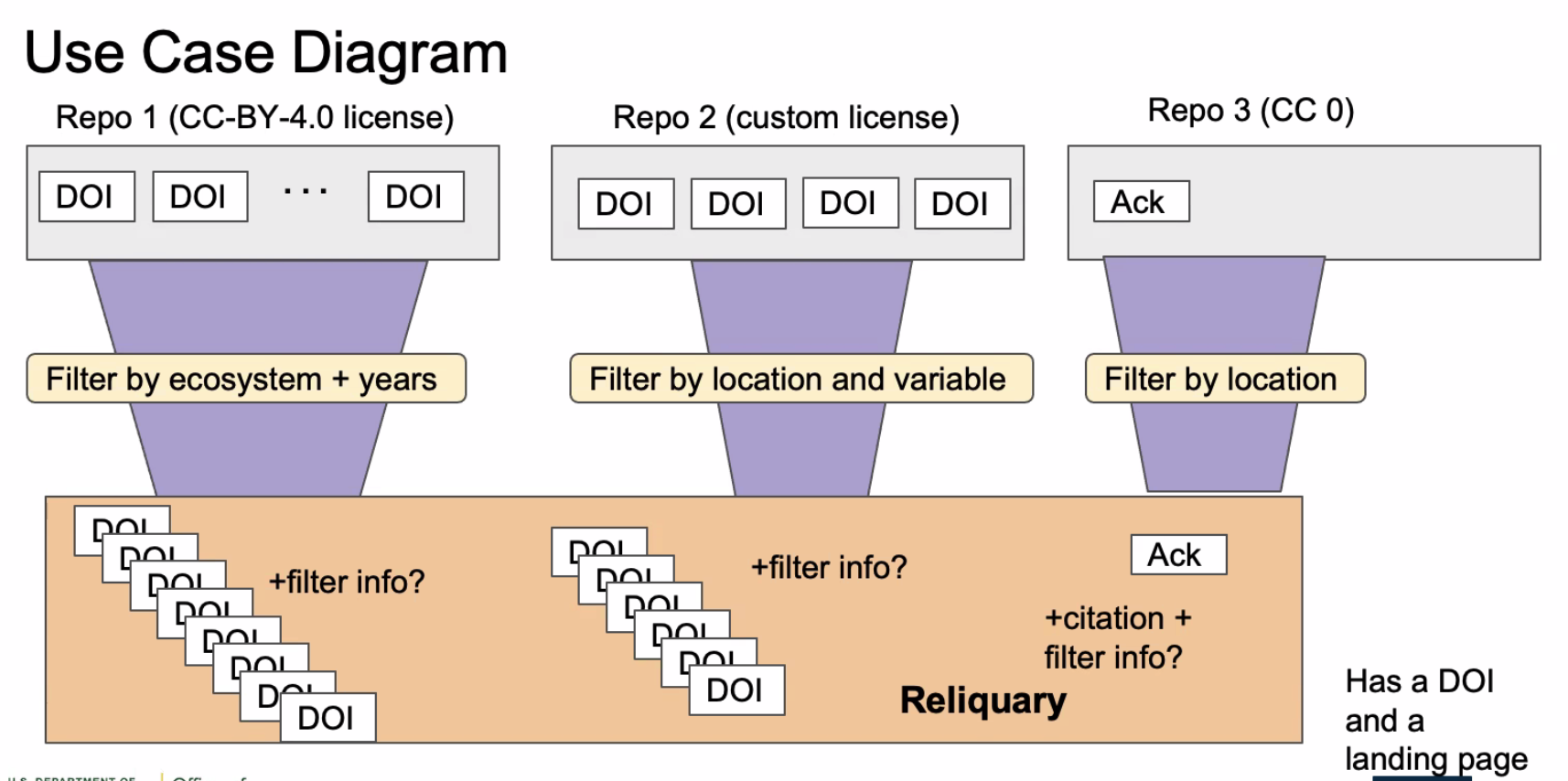
Working in context of large scale/collaborative projects, one level below IPCC example, contributing to many projects/spaces

Ameriflux - 500 contributing sites, independent research teams, put up site with group helping to gather data, been doing for years - first driving use case -> people couldn’t have 500 citations

Large teams, watershed, tropics, 70-100 subgroups gathering data, publishing data, 100-200 data packages a year, unique DOI to that data

Any one users could have used 200 datasets from Ameriflux, 100 from watershed, one from USGS, doing this in supplemental information/documenting

Queries for subsets of years, locations



Not assuming granularity at first, supplemental information in reliquary, annotation on/in reliquary

Sources require identification as well

Acknowledgements are also getting very long - DOE requires acknowledgement

Each user applying different filters, don’t control, might just want a list of sites

From Bruce Wilson to Everyone 05:09 PM

It occurs to me (and maybe I’m just slow), that to build on my example, the reliquary could provide the citation of both the general (the specific data products from which we created the subset) and the specific (the specific version of the algorithms by which we did the subletting and reproduction).

From Madison Langseth to Everyone 05:17 PM

It seems like the original use case was to deal with credit, so I would advocate for tackling the credit concept first.

**Discussion/Next Steps** (Moderator: Deb Agarwal/Shelley Stall)

Small group, definitions group - w/ broad application in mind

Need variety of cases on group

COMSES use case - citation for credit is important

Document, what information be (in Word doc), what level would credit be counted

Will this meet your needs

Requirements list (failures part of discussion as well, kick the tires) - would use RDA as way to collect requirements / float at ESIP as well / connect with data granularity group at RDA too

* The RO-Crate team were proposing trying out a PoC to figure how this would work with the use cases, just to kick the tyres on something concrete and feed into the requirements - through BY-COVID

Elisha Wood-Charlson -bring in e-notebooks concept as well

Start group after the AGU FM

GitHub as collaboration platform (suggested by Bruce)