Data Citation Community of Practice Workshop for Data Citation
Chat from 29 October 2021

11:02:16 From Laura Lyon, AGU to Everyone:
Hi everyone, agenda and notes can be found here: https://docs.google.com/document/d/1IbAVuwvqJh_xAiR9K9sXSwwX1rV6m59BkJNSJ0XiW0I/edit?usp=sharing
11:08:43 From Shelley Stall to Everyone:
If you haven't yet signed in, please add your name to the document.
11:10:15 From Reyna Jenkyns to Everyone:
Is the term accession here synonymous with 'reliquary' or collection?
11:10:58 From Maggie Hellström to Everyone:
@James: your curation at the point of acquisition?
11:12:50 From Stian Soiland-Reyes to Everyone:
what other types of PIDs are used here given that DOI is a type of PID?
11:12:55 From Reyna Jenkyns to Everyone:
I agree the non-hierarchical relationships are super important to be able to handle in whatever goes forward.
11:16:31 From Nancy Ritchey to Everyone:
I agree with the confusion on the term collection
11:16:34 From Carole Goble to Everyone:
Accession is used in the biosciences for submission
the big bioscience databases have "accession numbers" as identifiers
11:17:10 From Reyna Jenkyns to Everyone:
I'm also confused why DOIs aren't represented inside the collection.
11:17:18 From Shelley Stall to Everyone:
Thanks Carole…funny how that wasn't clear to me until just now.
11:19:02 From Maggie Hellström to Everyone:
BODC "collection" -> "thematic aggregates" ?!
11:19:49 From Mark Parsons to Everyone:
Is this approach unique to BODC or do other seadatanet archives do this too?
11:21:20 From Maggie Hellström to Everyone:
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?
11:23:14 From Shelley Stall to Everyone:
Hi everyone, agenda and notes can be found here: https://docs.google.com/document/d/1IbAVuwvqJh_xAiR9K9sXSwwX1rV6m59BkJNSJ0XiW0I/edit?usp=sharing
11:23:45 From Deb Agarwal to Everyone:
@Maggie - I think what allows them this complexity is that they have all the information and would build the reliquary for the user?
11:25:25 From Maggie Hellström to Everyone:
@Deb: maybe their "designated community" in OAIS-speak understands the complexity, but I worry about non-traditional end user communities!

11:25:46 From Reyna Jenkyns to Everyone:
We're doing camera data similarly - one DOI per deployment, and then identifiers for each file.
11:25:49 From Stian Soiland-Reyes to Everyone:
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)
11:30:21 From Nancy Ritchey to Everyone:
@Shelley the use of DOIs for the 'reliquary' and UUID/PID for the granules/files makes sense to me.
11:30:36 From Stian Soiland-Reyes to Everyone:
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.
11:32:12 From Stian Soiland-Reyes to Everyone:
we can encourage UUID use as base case, so at least they are globally unique compared to "dataset 42"
11:32:25 From Maggie Hellström to Everyone:
@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".
11:33:11 From Maggie Hellström to Everyone:
(Compare the Chinese (?) initiative to assign GURPIs to each individual packet of milk that is sold.)
11:34:04 From Stian Soiland-Reyes to Everyone:
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")
11:35:58 From Reyna Jenkyns to Everyone:
Looks like these handle IDs almost double as a query PID for the DOI that is pulled into the reliquary.
11:36:17 From Oliver Bandel to Everyone:
What about IPFS for circumnevt the URL problem?
11:39:01 From Madison Langseth to Everyone:
Did I hear correctly that the authors using the data would be responsible for creating the reliquary PID as opposed to the repository?
11:40:02 From Stian Soiland-Reyes to Everyone:
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.
11:45:06 From Maggie Hellström to Everyone:
But "anyone" can create their own collections and register these, e.g. at DataCite!
11:52:52 From Justin Buck to Everyone:

These discussions are really helping to shape how we need to show/communicate this at AGU
11:53:47 From Stian Soiland-Reyes to Everyone:
+1 Justin - we need to show both the needs AND the way forward without seeming like it's going in many different directions. And I get the feeling we are aligning well.
11:54:16 From Carole Goble to Everyone:
+1 not "here is my solution, what's your problem" :-)
11:56:09 From Caroline Coward to Everyone:
500 citations=consternation. 1 reliquary = relief.
11:56:54 From Mark Parsons to Everyone:
I'm beginning to think a reliquary is typically represented as a graph
11:57:15 From Shelley Stall to Everyone:
Linked content?
11:57:21 From Mark Parsons to Everyone:
yes
11:57:25 From Shelley Stall to Everyone:
Me too
11:57:39 From Mark Parsons to Everyone:
With nodes AND edges
11:57:58 From Shelley Stall to Everyone:
oh…interesting
11:58:00 From Shelley Stall to Everyone:
yes
11:58:14 From Stian Soiland-Reyes to Everyone:
Mark - agreed - it's a selection from the theoretically very large PID graph of possible citations
11:58:54 From Stian Soiland-Reyes to Everyone:
with a few glue-edges or proxy-nodes where needed (when the granularities didn't meet up)
11:59:29 From Mark Parsons to Everyone:
Hence the importance of edges (i.e verbs)
12:00:05 From Martina Stockhause to Everyone:
Agree with Mark. And the data granules in the reliquary have to include relations to other PIDs or DOIs they are built on.
12:00:09 From Maggie Hellström to Everyone:
The main problem I have with the "reliquary" concept being used for holding vessels for data (and other research-related digital objects) is that history of course has shown that many objects that were identified as "relics" and stored in actual physical reliquaries were unfortunately of questionable origins and/or "holiness"...
12:00:22 From Stian Soiland-Reyes to Everyone:
@Mark so it's not all "hasPart" ? Richer provenance on how selected/used?
12:00:44 From Mark Parsons to Everyone:
@stian correct

12:00:51 From Justin Buck to Everyone:
it helps to assemble the graph connecting the publication to the data sources (for transparency without the need to duplicate the data sources into new DOIs), part is only part of the broader graph in the long term
12:01:17 From Caroline Coward to Everyone:
@Mark you've been drafted into the small group.
12:03:04 From Stian Soiland-Reyes to Everyone:
@Maggie - we are hoping for a better name, but I think Shelley/Deb used "reliquary" as a working term! But perhaps there's something there.. otherwise boring indistinguishable data becomes marked as golden/holy just because it has been added to a reliquary - so it hopefully had some value.

It's basically the academic citation network (or PageRank algorithm at Google) again at a more granular level.
12:03:16 From Mark Parsons to Everyone:
Citation is not a goal in itself it is a means to a goal—transparency, credit, access, impact….. I think we're really talking provenance
12:03:35 From Caroline Coward to Everyone:
@Stian, @Justin, and @Martina are dangerously close to being drafted as well.
12:03:43 From Bruce Wilson to Everyone:
I think there's often a tension between citation for credit and citation for reproducibility.
12:03:54 From Mark Parsons to Everyone:
@Bruce Yes!
12:04:00 From Nancy Ritchey to Everyone:
@Bruce I agree
12:04:03 From Stian Soiland-Reyes to Everyone:
+1 Bruce - different granularities and access requirements!
12:04:06 From Maggie Hellström to Everyone:
@Bruce: interesting comment - can you expand on this?
12:04:07 From Caroline Coward to Everyone:
OMG Hi Bruce!! 👏
12:05:36 From Christine Laney to Everyone:
Great point Bruce
12:06:36 From Stian Soiland-Reyes to Everyone:
@Bruce same on versioning and mutability - more important to lock down for reproducibility. For credit it is traditional to still give credit even to old contributors (in fact it's a problem in regular citations of living resourcex in that new contributors are not credited because the established "reference citation" is old)
12:06:44 From Bruce Wilson to Everyone:
@Maggie — it also applies to things like subsets. We provide, as an example, subsets of some satellite data products. In terms of credit, we want users to cite the original data products. In terms of reproducibility, it makes sense to cite the very specific subset that we

created in response to the user's request. If the user cites our service, then the primary dataset doesn't get credit.
12:06:58 From Bruce Wilson to Everyone:
@Stian — concur.
12:07:17 From Martina Stockhause to Everyone:
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.
12:07:28 From James Ayliffe to Everyone:
Really sorry I have to go
Really good and interesting
12:08:24 From Mark Parsons to Everyone:
I think it's time to stop talking about citation and talk instead about the specific concerns
12:08:39 From Maggie Hellström to Everyone:
@Bruce: thanks for the example! I think we (as data producers) have to be more clear and explicit on how we expect end users to cite and/or refer to data sets (parts or as a whole). Maybe it's not enough to have associate only one citation string with e.g. DataCite records?
12:09:08 From George Porter to Everyone:
I have to run to another meeting. Thanks you all for trying to tackle a quite difficult but important set of problems.
12:09:34 From Nancy Ritchey to Everyone:
@Mark +1
12:09:52 From Bruce Wilson to Everyone:
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).
12:10:42 From Caroline Coward to Everyone:
I'm more sideways motion...
12:10:50 From Maggie Hellström to Everyone:
Which P18 breakout is that in?
12:11:40 From Howard Ratner to Everyone:
I like where this is headed but we just need to make sure the "reliquary" is easily discoverable in as many places as possible
12:12:10 From Mark Parsons to Everyone:
Parsons and Fox. 2014. Why Data Citation Misses the Point https://doi.org/10.5281/zenodo.1241521
12:12:49 From Stian Soiland-Reyes to Everyone:
quite a lot of citation/FAIR at AGU Fall meeting!
https://agu.confex.com/agu/fm21/meetingapp.cgi/Search/0?sort=Relevance&size=30&page=1&searchterm=citation
12:12:54 From Justin Buck to Everyone:

we have a presentation at AGU too
12:13:29 From Mark Parsons to Everyone:
I'm not sure we've fully defined the problem yet
12:13:31 From Shelley Stall to Everyone:
P18 Breakout 3
12:15:55 From Caroline Coward to Everyone:
There are 6 names on the working group list so far. We're nearing saturation, so let me know if you'd like to help define the thing.
12:15:59 From Mark Parsons to Everyone:
Earth science data has such use outside of academia which deserves credit
12:16:07 From Mark Parsons to Everyone:
such=much
12:17:49 From Madison Langseth to Everyone:
It seems like the original use case was to deal with credit, so I would advocate for tackling the credit concept first.
12:17:53 From Mark Parsons to Everyone:
+1 on requirements
12:18:51 From Elisha Wood-Charlson to Everyone:
This is great! My use cases fits in and happy to support/adopt. If anyone needs to include the concept of "e-notebooks", I could join.
12:20:15 From Bruce Wilson to Everyone:
Need tests that are off the happy path....
12:20:20 From Howard Ratner to Everyone:
+1 Caroline
12:23:20 From Martina Stockhause to Everyone:
Joint RDA/ESIP would be good in my view
12:23:48 From Bruce Wilson to Everyone:
Github repo as a collaborative platform?
12:24:17 From Chris Erdmann to Everyone:
The site is on GitHub :)
12:25:09 From Oliver Bandel to Everyone:
Alternatively GitLab
12:25:29 From Reyna Jenkyns to Everyone:
Also come to the Data Granularity WG session at the RDA Plenary next month! A lot of these discussions relate to what we are trying to pull together.
12:25:30 From Carole Goble to Everyone:
RO-Crate is github
12:26:01 From Elisha Wood-Charlson to Everyone:
+1 Reyna!
12:27:55 From Carole Goble to Everyone:
that is a great point maggie
12:28:09 From Elisha Wood-Charlson to Everyone:

BCO-DMO will be at OSM and can collect community data. If we have something standardized to ask, I can make sure they are aware of this. Don't see Danie on today.
12:28:19 From Mark Parsons to Everyone:
Dynamic Data Citation is also interested
12:28:21 From Carole Goble to Everyone:
yep - Stian and I are in the FDO forum
12:28:36 From Chris Erdmann to Everyone:
Gdocs has some issues in China, I've used Etherpad to share with colleagues in China
12:29:25 From Elisha Wood-Charlson to Everyone:
We have an RDA BoF session w/ Australia around microbiome data, but I can share these use cases w/ them as well.
12:29:27 From Stian Soiland-Reyes to Everyone:
In RO-Crate we do alternate times at 20:00 UTC and 08:00 UTC - always slightly awkward for everyone ;)