IMC VWC - Multi-repository landscape for archiving LTER data

March 12, 2018

Notes taken by Suzanne Remillard and Gastil ; compiled by Jason Downing.

Participants: Margaret O’Brien, Suzanne Remillard, Gastil, Marty Downs, Wade Sheldon, Stace, Stevan Earl, Kris, Yang Xia, James Conner, Don Henshaw, Jonathan Walsh, Sven Bohm, Renee Brown, Corinna Gries, Tim Whiteaker, Matt Jones, John Porter, Duane Costa, Mary Martin, Kristin Vanderbilt, Mark Servilla, Adam Sapp, Adam Shepherd, Chris, Li Kui, Hope Humphries, Ken Ramsey, Jason Downing.

Pre-meeting set-up:

Background: Since a prototype data catalog was added to the LTER website in the late 1990's (the LTER Data Table of Contents*), there has been single canonical location to find a list of core data sets archived by LTER sites. Even as we've collaborated as a network on development of data federation systems (KNB, SEEK, and now DataONE) we've always maintained a distinct identity with our own catalog, and our practices, standards and working-group activities have centered around that shared resource and vision. However, with the transition from a dedicated network office to a shared data center model (EDI), and the increasing expectation that sites deposit (or link) data to NSF-program-specific repositories like the Arctic Data Center (ADC) and Biological and Chemical Data Management Office (BCO-DMO), it's worth discussing whether this approach is still valid. And if not, what that would mean for LTER IMC goals and activities. Representatives of EDI, ADC and BCO-DMO will join us this month for an informal discussion that we hope will lead to more detailed and strategic discussions in the near future or at the ASM.

Just a reminder that our March Virtual Watercooler session today will focus on how the LTER IMC should adapt to the emerging multi-repository landscape for archiving LTER data. For example:

* what approaches are sites using now to accommodate multiple repositories?

* should LTER sites maintain a comprehensive catalog on their own website, or just point to external repositories for catalog support?

* what is the value of maintaining a "branded" LTER catalog in this new landscape?

* should we broaden our IM guidelines (and working groups) to accommodate additional metadata standards (e.g. /ISO 19115)?

VWC Discussion (moderated by Wade Sheldon):

Mandates and/or requests for LTER sites to work with other repositories (Artic sites with Artic Data Center; Ocean sites with BCO-DMO; ag sites). EDI advisory committee advised that EDI shouldn’t run dedicated portals, like for LTER. This causes logistical issues, branding issues, etc.

Explanation of various data centers:

Adam Shepherd (BCO-DMO) – biogeochemical data from cruises; images, models. Interaction with LTER has been harvesting EML records due to request from NSF officer, Dave Garrison.

Funded from NSF OCE (BIO and Chem). Primary program officer is Dave Garrison. Funded since 2006 to manage full data lifescyle. Marine biogeochemistry. Small tabular files from cruises. Recently image data nad model results. Manages “non-Core LTER” datasets coming out of OCE Bio and Chem. Interaction with LTER has been harvesting metadata records from D1 so OCE-funded research shows up at one repo. Potential repository for new LTER sites. New idea. Trying to understand how BCO-DMO can best serve LTER.

Matt Jones (ADC) – NSF funded repository from artic funded research (polar programs); they have an extremely progressive and proactive program. The quality of the data and metadata must be very good; the bar is very high. Took over management in 2016 from Arcatis in Boulder. Would like a metadata record for every dataset; they prefer the data, but if the data are in another repository, that is OK.

Ken Ramsey has mentioned that the LTARs are interested in a similar repository. Request for all LTAR data to go to the National Agricultural Center. Exposed a problem with harvesting from DataOne, but appeared to produce duplicates due to updates/edits of data and metadata. Some issues between metadata standard cross walking. (National Ag Library)

Seems like it is a desire for program managers to account for their program and make the data discoverable. It creates an interesting and unique situation about where to upload data.

Wade (GCE) – Mines DataOne to put into BCO-DMO. Listed as ‘data hosted in other repositories’ in BCO-DMO, based on keywords (I think titles for SBC) to populate those lists. Discussed using more structured keywording or the new EML 2.2 structure for funding metadata. Can show URL links as “data hosted in other catalogs”. Links resolve to EDI’s portal.

Our discussion could be do others cross-reference data and how do they deal with it?

Margaret

Slight variation on general pattern for LTER sites to cross-list in BCO-DMO. They use the scope to find datasets. We (SBC) also have some leveraged non-LTER data in BCO-DMO tightly coupled to our research. Our diff solution… Adam queries a specific project title from d1 to list datasets. Asking Adam how many requests have come to you like that Kelp Metapopulation one?

Adam

Eleven different variations of that D1 query.

What we (BCO-DMO) are doing there is to query project fields in datasets in d1 to capture all data Dave Garrison wants to see in a sustainable way so we do not have to go back and tweak that pull process from DataONE if we were to key in on a field like a datasets identifier - we want the SOLR query that will capture the grouping of data over time which frees up the LTER IM from coordinating with BCO-DMO everytime an update is needed. Finding best query to capture that. Run overnight. Queries D1 SOLR endpoint, creates Linked Data triples and a W3C provenance record to capture the query, the software version that executed it, the D1 result set on that day, and how the results are linked to BCO-DMO projects/awards.

Example:

https://www.bco-dmo.org/project/563139

THis is actually a collaborative project, with 3 NSF numbers. So the query runs for all three.

It queries eml:eml/dataset/project/title for the NSF project title.

Note for LTER IMS: This is a project that leverages SBC LTER, and our IM office handled the data curation. in the EML, we put the leveraged project (the kelp metapops project) as the primary project, and the SBC LTER project is a child project (relatedProject)

Here is the raw XML from pasta:

https://portal.edirepository.org/nis/metadataviewer?packageid=knb-lter-sbc.102.1&contentType=application/xml

Adam says: Each new query gets added on. Simple to add more. They have 11 running now.

Margaret

Tricky part was to find an EML field reliable that is indexed by D1. We used the project title.

Adam

DG cares most about checking in on an OCE award, so we fell the best "glue" between a grouping of LTER data and a BCO-DMO project is to use an award number if possible.

Matt: They try to map award numbers into funding section of EML. That is why this new structure was added to EML2.2. Meant to be relatively generic, but a place to put award numbers.

John P.: Have done very similar things to BCO-DMO; it serves as a good model. It is not good to have similar data in multiple systems. The way that DataOne does it is a good way to avoid redundancy. But there are other types of datasets, especially huge (multi-gigabyte) data that doesn't go into PASTA well. EML is good because it does cross walk well (like into ISO 115), but other standards don't map well back to something like ISO.

Jason: BNZ has lots of datasets that are in ADC. We had to upgrade our EML to include the project funding codes. We have some implementation on our end to track the cross-listed datasets, adding new stuff in EML… so they get updated and (replicated?). Matt has been working with me on this.

Matt

Should work great, replication approach in D1 to link these repos. Same ID at diff notes means not duplication problem. (?) Issues for this to work: whoever uploading, needs to set replication policy where it goes. Maybe not yet inside the current work flow.

Mark Servilla

Has been in place for awhile now. Was told ADC was not doing replication with other than ___, at least as of a month ago.

Matt

We have not yet turned on general replication. Have 20 Tb of data. Would automatically try to push all that out. Currently manually triggering each datasets to replicate across so as not to flood the whole network.

Mark S

You can drop the replication info/policy into system metadata can go into addl metadata in eml. That will push it into the target replicate site. Working with Jim L's site. Once proven will advertise more broadly.

Social/political issue - if we replicate data in pasta, feel compelled to support inbound replication in kind. Have space issue currently. Could support site-based repl. But not remote sensing dataset corpus.

Matt

D1 has node replication (thing). Code was written to allow codes to say where accept replicates from. Not sure currently deployed.

Mark S

This summer could test that.

Wade: What is the value of these expectations of having all data in one place? Do we want pointers?

John: One nice thing about PASTA and other services that are using web services, could point to multiple portals. The logical place to do some of these things are at DataOne. But the web services seem to be focused on the member nodes. It would be good to know what their capabilities to build these on top of DataOne.

Matt: Do-able. The API of D1 does give full access. The D1 portal uses that API. The documentation leaves a lot to be desired. We wrote up a simpler api doc for tehd 1 datacenter giving high level view. Do not need to know all those things to get started.

Tim Whiteaker: Solr API? It's part of it, but there are other APIs to do other things.

Matt

Other APIs are for reytrieving datasests, download counts, …

https://github.com/twhiteaker/Solr-JavaScript-Search-Client

https://arcticdata.io/catalog/#api

From John to Everyone: (12:38 PM)

https://releases.dataone.org/online/api-documentation-v2.0/apis/ITK_APIs.html

is the web page for the DatONE API

From Mark Servilla to Everyone: (12:36 PM)

here is the link to DataONE's APIs for both member node and coordinating nodes: https://releases.dataone.org/online/api-documentation-v2.0/#api-reference

From Tim Whiteaker to Everyone: (12:37 PM)

https://github.com/twhiteaker/Solr-JavaScript-Search-Client

Corinna: We (EDI) are looking for feedback. A strain on resources to maintain more than one portal. Starting do diverge. Have discussed that anyone can develop a portal to the data in pasta. We have had other groups request their own portal like LTER's. So far they realize that LTER has been paying for it.

If people are willing to expend resources, possible for anyone to build DIY.

Where all these data are going, seems should discuss different approach to pulling these data together. Such as DataONE. It should not matter where these data are. We seem to be scrambling every time a new program officer asks a new thing. Ask program officers to agree what should be done, where the money should go, how these should be used.

Shouldn't DataOne be the place to find all data? Why do we scramble with each new program officer who wants a specific repository used? We should probably have this conversation with NSF.

Matt: Yes, DataOne could play that role. They don't store the data, but they can point to repositories.

Corinna: Seems like the 'marketplace' issue is coming back into play. There are things that could be outsources and not build custom solutions.

Wade: We should push back on NSF about yet another repository so they could have a check off. This all adds work to people's plates.

Matt: the program officers are inflicting control over their own programs. They need to get universal agreement across programs.

Margaret: we need to get them to understand what is technically available right now.

Wade: Our activities past decade focused around pasta development, implementing lter data portal, related efforts ie units, congruence. IMC activities. Momentum. A lot of repos are not EML based. EML can be leveraged, but… new LTER sites new mindset. One arctic site has existing ISO 19115 approach. pro/con of broaden thinking, embrace different meta specs. Maybe not funnel all thru eml.

Ken: Federal agencies are required to create ISO in the absence of FDGC.

Matt: DataOne can create different standards.

Ken: Is there any effort at DataOne to corral these different metadata standards?

Matt: In general, if people are creating derived datasets, DataOne will link to the originals.

John: Do not want to go back to clearing houses. PASTA is good because the data are actually there. That is very powerful. A lot of the systems out there don't do that. It's good to have a guarantee that the data is actually out there.

Corinna: But that is why we need this ecosystem of smaller data repositories because the community can work together to create high quality metadata. We need to push back to NSF about replicating data in repositories, we need to replicate metadata only. There is a problem with the ag system because they do not point back to the original data.

Ken: LTER has their own identifiers. Nicole Kaplan is in LTAR so can advise from within. LTAR is open to communication, in contact with EDI, using published training material. Just early on. As a federal agency they can mandate how to do things. Tabular data, remote sensing, data collected for agricultural research.

Corinna: Yes, it's Cindy Parr that is running this and they know all this, they just haven't got very far.

Ken: Maybe we should try to connect to this group at ASM.

Wade: Yes, we need to explain alt identifiers.

Gastil: At a previous ASM, Dave Garrison and Cindy Chandler, SBC, MCR did get together to discuss an earlier concept of this. Between now and ASM, if we come up with concrete plans/request, ASM is a good place to formally discuss this.

Wade: Yes, a good place to take a view across data holdings. (kick the tires of the api's).

Gastil: program officers aren't aware of the technical possibilities; a demo could point out those that can be maintained between those that will fall apart.

Matt: another good place to discuss this is ESIP because it is broader than LTER and these are issues that this group has been grappling with.

Margaret: ESIP is good for the technical aspect, but to reach program officers, ASM would be a better venue.

Marty: There will be a good number of program officers at Science Council. We can plant the seeds.

Margaret: We need to evangelize with the program leaders (LTER PI's). Policy versus practicality. What is reasonable?

Marty: Yes, that was my point.

Matt: We could advocate for a policy position, ideally one that we all agree upon. There are benefits to replication for long-term preservation. Do it in a way that is maintainable. We should replicate and not duplicate. Ag is a good example of how NOT to manage these data.

Corinna: Does anyone want to put these things together and propose sessions at either ESIP or ASM?

Matt: I'll do one at ESIP.

Adam: I'll help Matt.

Corinna: Sure I'll help too.

Wade: we'll discuss with IMEXEC and identify champions to put together a program.

Chat Window Comments

From Renée to Everyone: 12:12 PM
Antarctic LTERs like MCM are now required to have data/metadata records in USAP-DC… mandated by NSF OPP.

From Margaret O'Brien to Everyone: 12:24 PM
This is actually a collaborative project, so the query has to run for 3 different project codes.
https://www.bco-dmo.org/project/563139

From Matt Jones to Everyone: 12:27 PM
DataONE provides a profile page for all member repositories: https://search.dataone.org/#profile/LTER

From Margaret O'Brien to Everyone: 12:29 PM
losing Matt

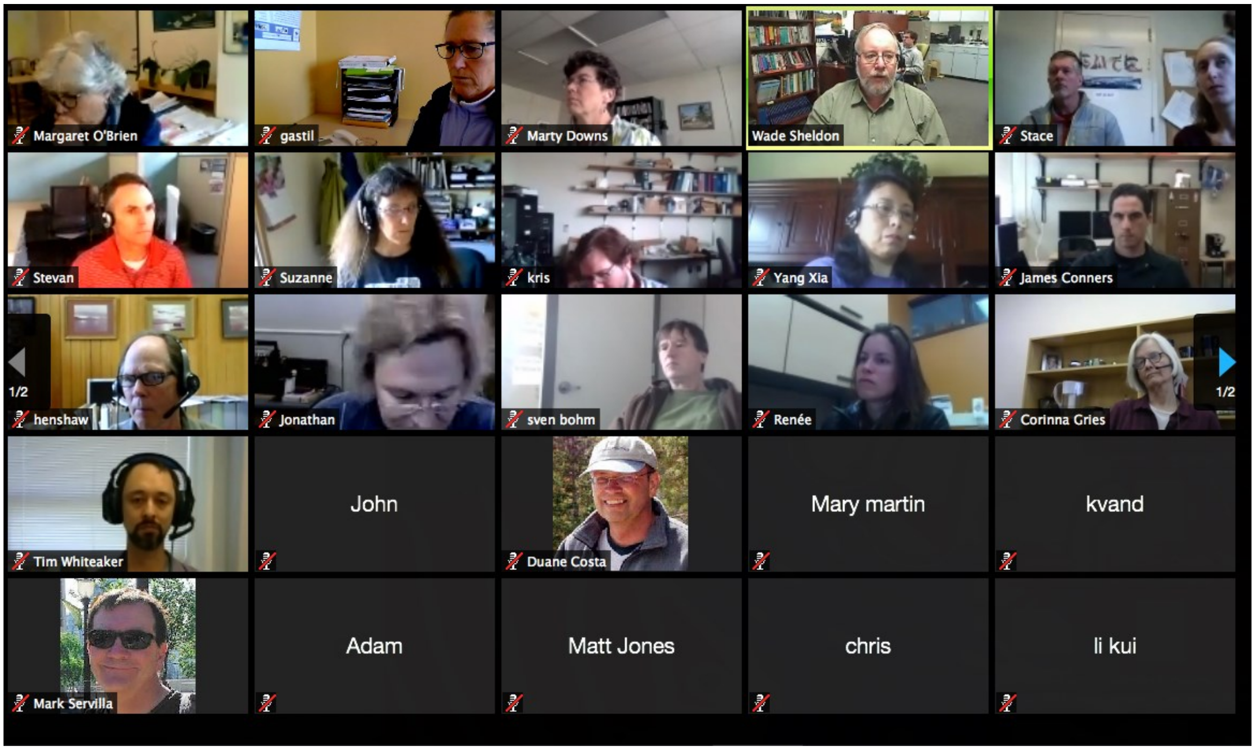
From Mark Servilla to Everyone: 12:35 PM
here is the link to DataONE's APIs for both member node and coordinating nodes:
https://releases.dataone.org/online/api-documentation-v2.0/#api-reference

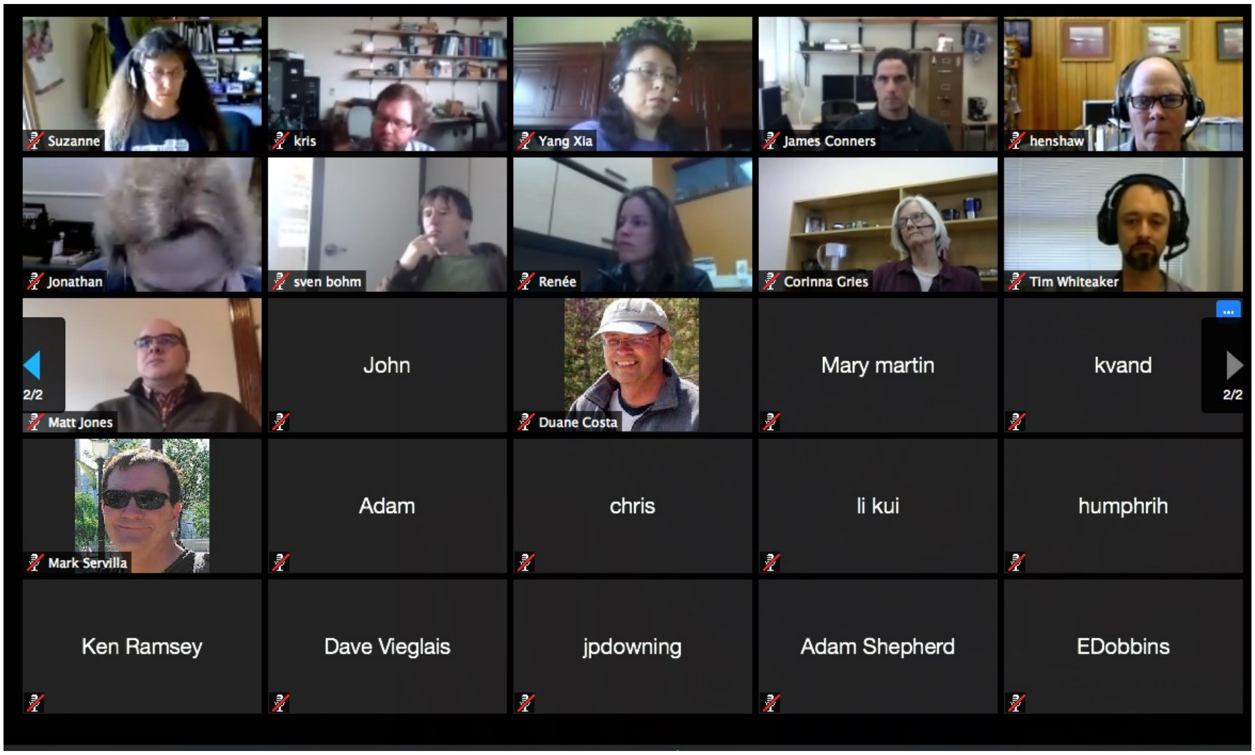
From Tim Whiteaker to Everyone: 12:36 PM
https://github.com/twhiteaker/Solr-JavaScript-Search-Client

From John to Everyone: 12:37 PM
https://releases.dataone.org/online/api-documentation-v2.0/apis/ITK_APIs.html is the web page for the DatONE API

From Matt Jones to Everyone: 12:38 PM
https://arcticdata.io/catalog/#api that's a simplified view of the API, focused on creating content

Attendees:


Margaret O'Brien
gastil
Marty Downs
Wade Sheldon
Stace
Stevan
Suzanne
kris
Yang Xia
James Conners
1/2
henshaw
Jonathan
sven bohm
Renée
Corinna Gries
1/2
Tim Whiteaker
John
Duane Costa
Mary martin
kvand
Mark Servilla
Adam
Matt Jones
chris
li kui


Suzanne
kris
Yang Xia
James Conners
henshaw
Jonathan
sven bohm
Renée
Corinna Gries
Tim Whiteaker
2/2
Matt Jones
John
Duane Costa
Mary martin
kvand
2/2
Mark Servilla
Adam
chris
li kui
humphrih
Ken Ramsey
Dave Vieglais
jpdowning
Adam Shepherd
EDobbins