Special Edition Virtual Updates Notes - 12/8 and 12/9 2008 Data Access at Sites and EB recommendations

Mon, 12/08/2008 - 4:17pm — sremillard [1]

Data Access at Sites and EB recommendations

Participants:
Monday (12/8): Corinna Gries (CAP; moderator), Don Henshaw (AND), Emery Boose (HFR), Hap Garritt (PIE), James Brunt (LNO), Jason Downing (BNZ), John Porter (VCR), Mason Kortz (CCE), James Conners (CCE), Ken Ramsey (JRN), Kristin Vanderbilt (SEV), Margaret O'Brien (SBC), Mark Servilla (LNO), Suzanne Remillard (AND), Jonathan Walsh (BES), Nicole Kaplan (SGS)
Tuesday (12/9): Corinna Gries (CAP; moderator), Wade Sheldon (GCE), Sven Bohm (KBS), Duane Costa (LNO), Barbara Benson (NTL), John Campbell (HBR), James Brunt (LNO), Linda Powell (FCE), Jim Laundre (ARC), Dave Balsiger (NTL), Theresa Valentine (AND)

Monday Notes:
Information Manager's Call - 12/8/2008
by John Porter
A. In attendance
1. Nicole
2. Hap
3. Suzanne
4. Emory
5. Corrina
6. Don
7. James Brunt
8. Jason
9. Mason
10. James (PAL)
11. Jonathan
12. Margaret
13. Ken Ramsey
14. Mark Servilla
B. Corrina was on committee that made recommendations...
1. based on Henry Gholz dis
C. Any real problems?
1. are we duplicating search capabilities by forcing implementation on each site
a) new enhancements to metacat looked great
b) but often titles are cryptic

2. Gholz is not complainging about finding data in MetaCat, but often CAN"T DOWNLOAD THE DATA
a) need to implement getting to the data.....
1) I looked at 5 or 6 sites
2) some sites can download directly from link in EML
a> others go to form
3) I thought Mark Servilla was working on this
b) but if link goes directly to data, lose ability to track
1) Data access server does allow this
2) is it operational?
a> will be in production by January
b> currently just prototype
c> document in SVN
1> changed considerably from RFC
2> urban.lternet.edu - NIPS, projects, VAS docs
c) don't think that is the issue here....
1) data just needs to be available from site's web site....
2) committee made multiple suggestions - fix on site and also fix in EML
a> But there is resistance to centralizing data access through EML
d) some sites DO want to use Metacat for providing access
e) some sites still want user registration or login to control and document access - but envisioned using Mark Servilla's system
1) some confusion about direct access to data (via login) or direct access with no login
2) auditing mechanisms are required by NSF....
f) talking about two different things
1) finding/locating data based on themes
a> better off supporting central solution....
2) downloading the data itself
3) better off doing centrally than 26 bandaids
g) single registration for multiple interfaces - eliminates sites needing to create interface for themselves....
D. Mark Servilla description of data download system
1. open HTTP access to data itself... See the document
2. every data URL made up of two parts - static and dynamic
a) replace the static part with data access component
b) join the two parts together with the site URL
3. how much effort will be required?
a) need one URL match
b) for each URL in EML need to be updated
c) need to add one entry in data access server to point to data
d) when you click on URL data comes back - no forms or files or query interfaces
1) set to text or Excel etc.
E. More discussion
1. Lots of problems PIs were having was that PIs could not negotiate to data, or would hit dead ends (not there, lost in jumble of other files)
a) but we should also use SiteDB to connect sites so that their datasets show up
b) similarly in ClimDB and HydroDB

c) would be good to have common interface with links to special datasets, etc.
2. I also like SITEDB, but it would be good to interface databases etc. and that will require more work
3. on download URL in EML needs to link to data in form described in EML document
4. email seems to have little to do with EML - Gholz et al could not find DATA collectecd by PIs in catalog
a) issue is data not showing up at all....
b) need to be able to find datasets under PI names
c) did not find "core/signature" datasetts
5. would be good to have siteDB be able to display the specified ways
a) and make that so it could be "included" in a site's home page as well
b) James Brunt - we are working on something along these lines - January rollout
6. need consistency within sites... are data represented consistently - what should reviewers expect to see...
7. data discovery page could be implemented centrally using EML
8. Data server requires us all to change our links.... but could we add field to EML for metacat link
a) can do that now - can have as many distribution trees as you need..... one to point to site, one to point to proxy server
b) would in some ways defeat purpose of proxy....
c) would be two links to the data....
d) could have both site specific and site, but require network version....
9. issues of SITEDB display on main and Intranet pages
a) already have stored queries within metacat
b) can't search within a site....
10. LTER data catalog - advanced search
a) but you can't pre-program the advanced search or pull out parts
b) also inconsistent application of core area names
11. requires rekeying data -for consistent core areas
a) would be good to do controlled vocabulary as well
12. Need consistent descriptions of all for core areas
a) need to see web site recommendations....
13. controlled vocabulary
a) GCMD keywords undergoing revision
F. Notes
1. asm workshops
a) involve scientists in workshop at ASM aimed at keywords
b) chain of responsibility for datasets
2. if go back to EB
a) three sites thought they would have problems implementing through Metacat download vs. data access server
b) who thinks implementing example data discovery page would pose a problem?
1) none
c) implementing data access server?
1) need more info on how that would work
2) need VTC (Suzanne will coodinate)

3. site centric solution deals with SiteDB
a) what does LNO need in terms of help to do this?
b) working on some changes already....
1) including more things in interface to SiteDB
c) will try do some demo in a month or so....
d) may want to make sure this is on the NISAC list of things to consider...

Notes #2 by Suzanne Remillard
How easy are the criteria developed to implement on each site? Can we share things across sites? The functionality we have on the Metacat may be duplicated if we implement this type of search across sites.
The data access problem has to do with access to the data (metadata is fine).
Is it the EML link that needs to go to the data? Or does it need to go to the site or a form to download data?
Are we talking about 2 difference issues: finding data or accessing/downloading data?
Finding data would suggest a central solution; downloading data may be a site specific solution.
Use SiteDB as a method to give these data access elements visibility. If every site has to implement these characteristics then it still may not happen. The interface can be site specific but yet still a network. This will likely require work on LNO to implement personnelDB and bibliographyDB.
Does the issue have to do with site PIs not having any LTER data in the catalog? This alert should help target those PIs to submit data. (making PIs accountable at a site).
NISAC is being charged with implementation of data access across the network not individual IMs.
It would be great/helpful to have access to criticisms from survey conducted by EB.
We should involve scientists with the controlled vocabulary (workshop at ASM); develop a chain of responsibility.
It really seems like SiteDB should be the place to access this type of information. It would be in a network view but with site specific links. Site's would have control of where things linked. This could be difficult to implement (one click to data) across sites because sites may have responsibility to other types of information (data access policy, etc.)
To implement changes to SiteDB, detailswould need to be flushed out.

Monday Notes:
by Emery Boose

How to implement or respond to recommendations from the Executive Board (EB)?

Corinna was on EB subcommittee that made recommendations. Phil Robertson passed some of these recommendations on to larger community. Process started by NSF's comments last summer. Corinna can return our comments to EB.

Implementation is probably not difficult for most sites. Is it a problem for any site? Can we help by sharing designs across sites?

Why should search capability in LTER Data Catalog be duplicated at site level?

Dataset titles are sometimes cryptic.

NSF is not concerned about finding datasets in Metacat. The concern is downloading the data itself. But data for some sites can be downloaded directly from the LTER Data Catalog (depending on site implementation).

Direct access to data via the LTER Metacat could compromise user tracking. The Data Access Server (DAS) will support data tracking and return user reports to sites. Designed to work with any access (EML, site website, publication, etc). In production by Jan 2009.

DAS design has changed considerably since RFP last year. Mark Servilla will circulate a link to the current implementation plan.

Major resistance in EB subcommittee to accessing data via LNO Metacat. Also a concern that some sites could not enable access via EML in time for reviews next summer.

Some sites have planned to use network search and download capability. Other sites have planned to implement these functions locally.

Auditing mechanisms have been requested by NSF.

Two different problems: (1) finding data, and (2) downloading data.

DAS will provide a common interface with single registration. But it allows individual sites to supply data separately.

Data URLs typically contain a static and a dynamic element.

Apparent consensus toward using central capability via LTER Metacat. Data link needs to be a file (text best), not a web page.

Customize LTER Network databases (e.g., ClimDB) to return a site's data. It doesn't make sense to develop 26 separate solutions. Use SiteDB as much as possible.

Centralized data discovery and access tools present a unified network. But tools are also needed at the site level for reviews.

EB recommendations also point to a need for capability / consistency at the site level.

Add a new field to EML for Metacat link? EML currently allows as many distribution trees as desired. Links to site, central proxy server, etc. But that might compromise central auditing, etc.

Site-specific URL could also be provided by Metacat.

SiteDB has a bland home page at present.

LNO Metacat supports site search using a stored query. It doesn't currently support query by site and an additional term. This combination might help address NSF's concern.

LTER Data Catalog is currently quite capable. Query page is implemented in JavaScript.

Develop SiteDB to address EB recommendations. Implementation of LTER core areas may vary from site to site. But search by core area must be supported.

Sites will need to re-key datasets for core area. We need to agree on how to do this. It would be a shame not to use this opportunity to incorporate other new controlled keywords.

Link from site home page should be "Data" (see earlier IM recommendation).

Corinna can forward specific comments to individual sites.

The controlled keyword list is not yet completed. Also no mechanism at present for updating the list. Need to develop tools to ensure consistent choice of new keywords.

Link to global change master directory list. Currently offline pending major revisions.

At least a couple of months (after list is completed) before tools will be available.

Need to schedule a VTC for controlled vocabulary group in Jan 2009.

Involve scientists in controlled vocabulary project. Perhaps a workshop at ASM?

ASM workshop on chain of responsibility for datasets.

Need more information on centralized solutions. Mark Servilla could offer a dedicated VTC.

User audit, controlled vocab, etc are problems that could be addressed in SiteDB.

What assistance does LNO need from larger group to develop SiteDB? First develop requirements, then consider backend solutions. Some upgrades already planned. Need to work out cross-links to Metacat, etc.

SiteDB demonstration in next month or so.

Site implementation of search capability, etc provides redundancy in case LNO server is down.

SiteDB is not currently on list of projects for NISAC. Move onto NISAC agenda. Not a major task, but needs to be accounted for by steering committee.

Tuesday Notes by Duane Costa

VTC Notes for December 9, 2008
Participants: Wade Sheldon, Sven Bohm, Duane Costa, Corinna Gries, Barbar Benson, John Campbel, James Brunt, Linda Powell, Jim Laundre, Dave Balsiger, Theresa Valentine,

Barbara -- We need a better definition between a "signature data set" and a "core data set".

Corinna -- Phil doesn't want us to spend six months working on these definitions.

Wade -- Re-organize the search page to provide a more consistent representation. It should be shared across sites. Search locally and point to the LTER Data Catalog.

Corinna – Issues and concerns of the EB:

1. Sites would have to implement data download via EML.
2. Problems about centralization, relying on a single LNO server which could go down.

We do agree that a centralized solution is preferable. Is there anyone here that has a problem with EML data downloads? (No hands were raised.)

Linda – We spent a year doing the data section at our site. Do all sites need to implement the same look & feel, the same interface? We have graphical icons. The icons go to our Table of Contents for data search and download, or they go to the Metacat. Will this need to be changed?

Barbara – Two things:

1. One-stop shop at network portal.
2. Allow sites and individual page to keep their functionality if it’s been working well.

A lot of sites have put a lot of energy into developing solutions that go beyond the Metacat. They can skin the Metacat if they choose.

Corinna – People have been complaining about too many clicks to get to the data. Going to Metacat must be made clear on the site's home page. People coming to your site want your data in a standardized format.

Linda – Our data is searchable by theme, long-term/short-term, etc. It offers more than the Metacat. PIs should take some responsibility for finding information. I did a quick survey and the results were that it was easy to find the data at our site.

John – It's not an either/or case of local versus network.

Wade – We need to be clear: here is a standardized network-level view, and you can also have your own local customized catalog. Here is a path to find the network-centric view of our data. We don't need to give up our own work. But we have to be careful how we implement this. We should avoid a "Chinese take-out" menu approach--choose one item from A, two items from B. We need an ongoing dialog. Some things will work, some won't. Individual sites will continue to provide a better window to their data. We shouldn't lose that.

Corinna – Standardization through the Site DB. How will this be implemented at each site?

Barbara – Do you mean Site DB or Metacat?

Corinna – Site DB.

James Brunt – Drawing on the data from Metacat.

Corinna – Nobody really implements that at their site yet.

Wade – Right, but we should push for it. Use links back to the LNO Metacat to find data from other sites.

Jim Laundre – We might be confusing Site DB as a standardized web interface versus accessing data. We should be focusing on a standardized data access web page instead of on Site DB at this point.

Corinna – Our original response to Henry's request for standardization of web sites was "Site DB". Data links should be found on the first page, and the same functions should be available at each site etc.. Now our response again is to do it centralized not at each site.

Jim Laundre – Henry's frustration is more with data access than with web sites. How do we create one-stop shopping for data access?

Corinna – How do we mesh the two approaches, site specific and centralized interface?

Jim Laundre – Are we looking at doing value-added, or using the same implementation across all the sites? How should we interpret Henry's request?

James Brunt – The EB has interpreted Henry's request already and issued a mandate.

Barbara – PIs at my site expect to have input on this at the next SC meeting.

Corinna – The next SC meeting is later than the first site review.

Wade – We need a multi-stage process. I don't want this to capture all the energy for data searching and access. Start with some derivative of the network Data Catalog as a first step. Use a drop-in query form, or some code that can be shared across sites. The expectation is that we are providing some standardization on what we're hosting at the sites.

James Brunt – A gradual shift to using a more standardized mechanism, without limiting the sites.

Jim Laundre – I'm curious to see what problems crop up on Metacat search. Even though we're all using EML, there are still things we are doing differently in the EML.

Wade – Controlled vocabulary becomes important. Although many parts of the search interface are standardized (for example, spatial search), the parts that involve topical searching are not standardized because we lack a standard vocabulary; our products are not key-worded appropriately. What is a core area? That's where Metacat breaks down.

Corinna – Data access is where it breaks down.

Barbara – Because data just aren't on-line?

Corinna – Yes, that's part of the problem.

Wade – Another problem is the staged roll-out of EML support. Some sites haven't fully described their tables and attributes in the EML yet because it takes effort. The staged roll-out is one part of the problem, another problem is just getting people to submit their data.

Corinna – Also, there are complaints that when the data is downloaded it is incomplete or of poor quality.

Duane – Could Wade comment about whether he sees the proposed Data Access Server (DAS) as part of the solution?

Wade – The DAS doesn't really address many of these problems, but one place where it does fit is that for sites that are not providing direct access to their data in their EML, the DAS would remove an impediment for them.

Corinna – Another area is that it provides centralized data access auditing which is important to implement the centralized data download.

Duane – And with centralized auditing, we can generate detailed data access reports to send to NSF. Imagine being able to document to NSF that there have been several thousand accesses of LTER data over the past year, for example, and to be able to break it down by site or by individual data sets. This could be a very powerful tool for demonstrating to NSF that LTER data is successfully being accessed.

Corinna – There seems to be resistance to implementing a standardized interface. It seems this will only happen if sites can point to the centralized system but still keep their local solutions in place.

Jim Laundre – A familiar interface across the sites is important.

Wade – There is resistance to giving up the specialized interfaces at the sites. Sites want to maintain a duality. The centralized interface should be complementary; they don't want to give up their specialized interface.

Barbara – There has been a lot of investment at the sites. Why should a more powerful system developed at the site be replaced with the lowest common denominator that is less powerful?

Wade – This fits in with centralized searching and controlled vocabulary. We don't want to invest time in a dead end. We should extract components that work for a short-term solution that can eventually be used in a long-term solution.

Corinna – Charge NISAC with evaluating the next steps.

Wade – NISAC will be meeting this Friday.

Attachment	Size
LTER-im LTER Lead PIs - Data Management - IMPORTANT.html [2]	5.05 KB
EB Data access recommendations Nov 2008.doc [3]	56.5 KB
LTER Data Page Elements Example.html [4]	4 KB

Virtual Updates [5]

Source URL: http://im.lternet.edu/node/384

Links:
[1] http://im.lternet.edu/user/78
[2] http://im.lternet.edu/sites/im.lternet.edu/files/LTER-im LTER Lead PIs - Data Management - IMPORTANT.html
[3] http://im.lternet.edu/sites/im.lternet.edu/files/EB Data access recommendations Nov 2008.doc
[4] http://im.lternet.edu/sites/im.lternet.edu/files/LTER Data Page Elements Example.html
[5] http://im.lternet.edu/taxonomy/term/169