Published on *LTER Information Management* (http://im.lternet.edu)

Home > IM Exec > IM Exec Winter meeting 2012-04-3/4 (Albuquerque, NM)

IM Exec Winter meeting 2012-04-3/4 (Albuquerque, NM)

Wed, 03/21/2012 - 12:00pm — mobrien [1]

IM Exec members in attendance:
Don Henshaw (AND, IMC Co-Chair), Margaret O'Brien (SBC, IMC Co-Chair), Sven Bohm (KBS), Emery Boose (Executive Board Rep, HFR), John Chamblee (CWT), Jason Downing (BNZ), and ex-officio members James Brunt (LNO CIO) and Yang Xia (LNO IM).

Non IM Exec members attended at various point in the meeting, including:
Scott Collins (EB Co-Chair, SEV), Peter McCartney (NSF Program Officer) via vtc, John Porter (NISAC Co-Chair, VCR) via vtc, Inigo San Gil (MCM IM), Saran Twombly (NSF Program Officer), Bob Waide (LNO ED), Kristin Vanderbilt (SEV IM)

AGENDA

Tuesday, April 3 -- all times tentative

8:30 Agenda review

900 - 1100
Meet with Scott, Bob, John Porter (NISAC co-chair) via VTC
- Plans for improving the network data portal to make available site data more visible
- Develop action plan for addressing perception that data are not available
- Consider other actions to put more data online or improve access
- Current available methods for assessing data packages

1015 - 1030 Break

1100 - 1200
IMExec business
- annual report from IMC to the EB
- upcoming rotations: co-chair, imexec, nisac, databits

1200 - 1300 Lunch Break

1300 - 1300
Update from active IMC working groups: http://im.lternet.edu/projects (liaison will have posted a paragraph on the wg/s page)
- Controlled Vocabulary - Don

- Units - Sven
- IMC Governance - Emery
- Web Services & Database Redesign - Sven
- GIS/Spatial - Jason
- Data package metrics - Margaret
- IMC Website - John
- Drupal Ecological IMS (DEIMS) - Kristin

1500 - 1515 Break

1515 - 1700
IMC 2012 one-day meeting (9 Sep) planning:
- identify IMC-only business, start a 9 Sep agenda
- Water coolers: pre-ASM updates and schedule for April-September

Wednesday, April 4 (with Saran Twombly, NSF) -- all times are TBA

0830 -1100
Introductions, agenda review, questions & update from Saran
- discussion points
-- bringing site level IM projects into production
-- Use of recent IM supplements to enhance data contributions from sites to the NIS.

1100 - 1200
VTC with NSF's LTER Working Group (TBA)
- 30 year review and responses
- developing site-IM priorities given network responsibilities and current funding
- Synthesis data project

1200 - 1300 Lunch Break

1300 NIS production working group proposals
- Upcoming workshops anticipated, based on 2011, 2012
- see http://intranet.lternet.edu/working-groups?title=&type=117 (some 2012 workshops not yet included.)

1400 - 1500
ASM
- IMC 1-day meeting on 9 Sep
- decide on our theme. consider likelihood of non-LTER participants.
- plan schedule; consider time that IMC might need in closed-meeting.

- ASM: main meeting, 10-13 Sep:
- potential IMC-led workshops to be proposed for the ASM
- ASM theme: "...LTER Network in the Anthropocene: Collaborative Science Across Scales"

1530 - 1600
White paper
- title: "Information Management in the LTER" (as described our 30-yr review comments)
- solicit writers, set a timeline

1600 - 1700
Wrap up, action items, assign tasks

NOTES

Data Availability, Data Accessibility Action Plans, Data availability metrics (Attended by IMExec members plus Porter, Waide, Collins, Vanderbilt, San Gil)

The meeting began with a brief discussion about whether or not a design charrette was still necessary to address the immediately known issues with the NSF Data Portal. John Porter informed us that these issues had been addressed through a Tiger team meeting, since it had been necessary to resolve the issue in late January / early February.

Bob Waide opened the broader discussion on data availability. He noted that there is concern within NSF that LTER data are not available and both he and Scott emphasized that the concerns are serious enough that they must be addressed. Bob categorized two kinds of data availability for LTER – data available through a site portal and data available through a network portal. John Porter, Bob Waide and Margaret O'Brien discussed a second division – machine versus human readable data. Some data require human interpretation and judgment, both to determine their suitability for a particular purpose and their content. Other data sets, however, are prepared in such a way as to allow programmatic ingestion into a work-flow, once it has been determined that the data are suitable for that work-flow.

While Bob and Scott acknowledged that these are legitimate divisions, the broader concern is that the LTER has created a moving target for data availability and that by continuing to revise and improve community standards and by introducing new technologies (e.g. PASTA), we are constantly chasing a moving target without fulfilling basic obligations. The general view at NSF is that basic descriptions that include discovery-level metadata and a basic description of the data table should be sufficient documentation and that many of the standards being pursued in the development of PASTA may be impeding basic data delivery. In response, the NSF has made the point that they would look to see improvements in data availability as soon as possible, with the aim being that, once the data are available with basic metadata, the LTER and the NSF can focus on science and move forward from the focus on data availability.

To that end, Bob Waide and Scott Collins have proposed two solutions:

1. LTER IMs should concentrate on making basic data and metadata available, both at their sites and in Metacat through a concentrated effort to improve their metadata in five key areas (see end of this document)
2. LNO presently has $170,000-$200,000 dollars available to put towards solutions that increase data availability at the site and network level. Their suggestion was to provide the funds to six sites that were very close to having a large number of additional data sets on-line so that the numbers of data sets across the network could rapidly increase.

IM Exec members and participating invitees responded with a series of additional suggestions to improve data availability and the perception of data availability.

1. Direct resources to improving already contributed site EML and/or make changes to LNO Metacat to improve visibility of existing data

Discussion: Margaret O'Brien and John Porter both pointed out that some data appear unavailable when in fact they are, but cannot be easily reached. Many of the problems Bob Robbins encountered in the video are a result of such problems. Margaret O'Brien noted that she and M. Gastil-Buhl now have a straw-man series of metrics for specific, frequently occurring EML constructs that create the impression that data are not available. Usually the cause is non-regular

EML implementation and the use of default EML stylesheets in the network's Metacat. She noted that addressing these two issues would alleviate a large part of this perception. Using checks and metrics to identify specific problems with EML would provide avenues by which systemic errors could be programmatically fixed. Doing so would increase the availability of data through the Network Portal by making the EML supporting those data consistent. She also noted that continued one-on-one mentoring on EML would help solve some sites' EML problems and facilitate additional consistent development of metadata, again potentially increasing data availability.

2. Direct resources to improve emerging information systems to improve efficiency and consistency in creation of metadata

Discussion: John Chamblee, Kristin Vanderbilt and Inigo San Gil discussed possible approaches to increasing availability centered on the additional development of database-driven data set documentation and structured metadata generation. Two potential projects would involve funding existing cross-site IM collaborations centered on DEIMS (the Drupal Ecological Metadata System), currently being used by eight sites, and the Metabase, currently being used by 4 sites. The idea is that, since these database systems can generate EML through views and application layers, the addition of more sophisticated tools for putting metadata into these systems would increase the pace at which IMs can prepare data.. Both of these systems are becoming widely used in the network.

A major concern for any of these approaches is that, while it is agreed by all that each of these solutions will provide long-term dividends for increasing data availability; none will immediately increase the number of data sets currently on-line. Any site agreeing to participate in projects supported by these funds would be strongly encouraged to demonstrate their readiness to increase data availability. But, even with such pledges, the timeline for increasing data availability may be longer than is desirable. The committee also considered the possibility of building site-specific scripts to standardize existing EML, but this option needs more research to be sure it does not result in ancillary problems of maintaining such data sets after the initial standardization.

Other potential solutions were proposed, including a push to update the NIN data in Metacat, further improvements to the LTER Data Portal to increase ease-of-use, and thereby the perception of availability, or a network-wide push to increase the availability of certain kinds of data. IM Exec was generally of the view that any of the proposed solutions above would alleviate some of the issues, but that LNO and IM Exec should solicit the opinion of NISAC by presenting a summary of the problem along with mini-proposals for potential solutions.

Several ancillary issues related to data availability were also discussed. IM Exec members noted that there seems to be some conflation of the NIS and PASTA and that concerns over PASTA, which is a new technology, are being applied to the Network Information System, which is a modular, distributed database consisting of Metacat, PersonnelDB, SiteDB, ClimDB/HydroDB, the LTER Bibliographic Catalog, etc, which have been in use for many years. Another concern is the possibility of a general perception that PASTA will replace site-level information management systems, which all site IMs and LNO agree that it was not designed to do. It was also noted that some NSF personnel may be concerned that the NIS will become an "empty archive," or a large development product with no data in it. IM Exec members were also concerned that many people including some in LNO, were not aware that some data sets will never be inserted into PASTA, which is currently capable of ingesting only delimited data tables, and that a plan for other data types is not available. A final issue of note was the potential barrier to accessibility represented by the login forms site that both site and network portals require before allowing the download data. The consensus among the attendees was that these registration requirements should be

voluntary or removed entirely (depending on the preference of the sites) and that we should clarify with NSF that we can track usage through other means (e.g. server logs). Note: NSF encouragement to eliminate login forms and other any barriers to data access was confirmed by Saran Twombly on Day 2 of IMExec.

IM Exec and NISAC Rotations and IM Exec Report to the Executive Board

Four members of IM Exec will be rotating off of the committee this year, Sven Bohm, Emery Boose and both IMC co-chairs. Emery Boose will remain as an ex-officio member until 2014 through his work as Executive Board Representative. There is consensus at IM Exec that continuity should be maintained in the Co-Chair position. In order to address this issue, one Co-Chair, Margaret O'Brien, has agreed to stand for a one-year term, in order to provide additional continuity and establish staggered terms of office for the future. Since such actions are not covered by the Terms of Reference, this will require a one-time exception to be allowed. This exception must be a formal action by the full IMC, to be proposed, seconded and passed by simple majority vote during the next IMC meeting. This request will be provided in writing prior to the IMC and included on the agenda.

Moving forward, IM Exec will begin considering nominees and will put forth a call for nominations in time for a list of candidates to be assembled and published ahead of the IMC, along with any statements of intent the candidates may wish to provide. Nominations from the floor for IM Exec may also be accepted. No NISAC terms expire this year. The members of IM Exec also discussed the need to seek compensation for IM Co-chairs, given the range of responsibilities they currently have and the additional workload that may be involved in leading data availability efforts. These discussions were raised with Bob Waide during the meeting.

The IMC co-chairs will produce the Annual Report to the EB, linking the report to Strategic and Implementation Plan objectives, as was the case last year.

Working Group Reports

Controlled Vocabulary (Don Henshaw)
The controlled vocabulary web site and web services (see http://vocab.lternet.edu/vocab/vocab/index.php) are working well using the TemaTres tool, which describes preferred and related terms and presents a hierarchic structure of terms.. Metacat search capabilities are also present as is a keyword distiller (http://vocab.lternet.edu/keywordDistiller/). The controlled vocabulary list has been translated into German and Chinese. John Porter is considering a proposal to improve the interface and increase the service's adoption at sites. This may be a summer water cooler topic. IMC members are visiting China this summer for an international vocabulary workshop, where our work will be presented.

Unit Dictionary (Sven Bohm)
The unit dictionary website now supports queries and additions (http://unit.lternet.edu/unitregistry/) and the web services are fully functional and in use at many sites. Most or all sites have submitted units. Linda Powell is the current chair but has asked to be replaced.

Governance (Emery Boose)
This group has been inactive this year, following the adoption of the IMC Terms of Reference.

Web Services and NIS Database Redesign (Sven Bohm)
The PersonnelDB Buy-out Proposal to add RESTful web services to PersonnelDB was funded. Sven, John, Wade, and James need to get together, make sure all existing code is in subversion,

add a JQuery plug in, work on style sheets, work with LNO to determine how PersonnelDB will work with PASTA's ID management services, and obtain a dump of the current contents of PersonnelDB from LNO. After the LNO site review, LNO may hire a contractor to do this work, rather than having James do it. The buy out time for James would instead go to supervision (perhaps to both James and John), rather than actual coding. There may be a prototype of some of the items above at ASM.

GIS / Spatial (Jason Downing)

The LterMAPS website is in the works and a best practices for converting GIS metadata to EML is underway. There have been challenges in this area related to the migration from ArcGIS 9.3 to 10, but these, and their solutions, are documented in the new best practices. There are several working groups being proposed for the ASM and the group is planning to set up and complete a virtual server at LNO. The GIS working group will participate in SiteDB group activities in LNO, scheduled for after the last day of this IM Exec Meeting. A Science Council presentation is planned.

IMC Website (John Chamblee)

This group is currently inactive, though Corinna Gries (NTL) continues to manage the website.

Data Package Metrics Working Group (Margaret O'Brien)

This group will provide a 5-10 page report on version 1.0 checks. The report will be comprised of tables and a summary of checker behavior. In order to fully understand the checks needed for data congruency, it was necessary to decompose some checks into the most atomic constituent elements. This means that, although there are a large number of checks, many are quite simple. There are 12 checks that are absolutely required for inclusion in PASTA, which are now implemented. The report will delineate those and include priorities for the remaining checks.. Sven Bohm has also implemented a web interface to run checks in evaluate-mode at KBS. The URL is http://26sites.org/ . Sites can also check data sets using a cURL command. There will be water cooler scheduled to report on the workshop in late spring or early summer. The related issue was mentioned: the imminent release of Metacat2 will need to be introduced soon, as this may create the illusion that the checker is done because Metacat2 is more stringent in its requirements than Metacat1. A Metacat 2.0 development copy is installed at LNO and the developers have initiated a comment period, seeking feedback during the late spring.

DEIMS (Kristin Vanderbilt)

The Drupal Ecological Information Management System (DEIMS) has, since beginning in 2008, expanded to include eight site participating in development. Several sites are now live and two international LTERs have adopted the system as well. The DEIMS group has funding for a training session this summer, but that agenda is not yet set. The group is also actively seeking funds for a developer. The SEV will hire a developer for some tasksing, including migration to Drupal 7 and the creation of a distribution package. A demo for ASM is under consideration. Developers are necessary, because of the size and complexity of the cod base. Some DEIMS participants attended Drupalcon and learned that the release of Drupal 8 is already being discussed and that this new verions will include integrated support for mobile browing interfaces.

The One-day IMC Meeting at the All Scientists Meeting

IM Exec determined that this year's annual meeting should be split into a closed morning session and an open afternoon session. The ASM on-line agenda now reflects this decision. The closed portion of the meeting will include our usual business meeting, elections, presentation of EML metrics as they apply to sites across the network, as well as discussions of current and ongoing plans to address overall concerns about data availability. The focus of the afternoon session is yet to be determined. Ideas include general topics of data availability, developing a process to

advance stalled projects (eg, UnitsDB, controlled vocab), or broadly themed projects (DEIMs, Metabase, EML mentorship programs, and SensorNIS). The themes from the one day IMC will be echoed throughout the ASM in other workshops and working groups and IM Exec will solicit working groups and help produce brief advertisements.

Meeting with NSF Program Officer, Dr. Saran Twombly (attended by IM Exec members and Kristen Vanderbilt)

IM Exec had a wide-ranging discussion with Dr. Twombly. The NSF Working Group for the LTER has been considering several recent developments related to LTER Information Management and considering their impacts. These developments include the contents of the Thirty-Year Review and the associated Robbins report (but stressed that this was not part of the Review), the arrangement through which LTER is to host data from Macrosystems Ecology and LINX awards, the NSF Data Management Plan requirement, and the LTER Network's development of the PASTA framework within the NIS. In addition, there is a perception across the NSF that LTER data are not as available as they should be. In response to these developments, NSF wants to clarify their understanding of the role that site-based Information Management plays within the LTER as well as clarify requirements for data availability.

NSF's position with regard to data availability is that data be on line and supported by enough basic metadata that those who find the data are able to tell what the data are about. That is the end of the NSF requirement at this time. To that end, NSF has already put significant funds into Dryad (http://www.datadryad.org), a system for submitting the data associated with a given peer-reviewed publication. Over time, Dr. Twombly anticipates that NSF will further encourage researchers to turn their data into data products and data papers. The GEO Directorate has already released a "Dear Colleague" letter encouraging such efforts. However, at present, and for most uses, the NSF views as sufficient a system that provides basic access because, as they understand the situation, it is the case that most researchers are not looking for sophisticated data products or data delivery systems, but instead are looking for basic data and metadata. For the LTER, the two main concerns should be providing these basic services and, perhaps more importantly, make sure that our long-term data – those data that represent the unique long-term efforts of LTER – are fully available.

Right now, the NSF believes that it is crucial we address the data availability decisively. Their goal is to turn discussions at the NSF away from LTER data and back to the more important issue of LTER science. The fact that we are still talking about data availability and that it remains a problem is the primary concern at NSF at this time. Related to this concern from the NSF are five ancillary concerns related to the way that the LTER network is currently pursuing data management:

1. We still need a viable and effective "one-stop shop" that people can use to search for LTER data. While researchers do need to find data at sites, they also need to find it at a central location. This tool needs to be available before the timeline for PASTA would make such on option viable.
2. The PASTA framework itself raises the concern that Information Managers and the LNO are regularly "raising the bar" for data availability, thus making the discussion about data availability an on-going topic and preventing release of the basic data and metadata that NSF is seeking.
3. There is some concern that PASTA will be an "empty archive," that is to say, one of many data repositories in existence that, because they are not tied to specific use scenarios, do not contain much data.
4. Long-term data are crucial to the LTER, but, because long-term data are also often old data, there is a concern that some of these data will be lost in the migration to PASTA because the metadata may not be present to support PASTA and because, in general, old or legacy data are much harder to work with than new data.

5. There is general concern that because PASTA was designed as a cutting-edge research project, it may not serve well as a basic data archive. Related to this concern are the resources being allocated to PASTA's potential as a driver of synthesis-related research, given that the NSF BIO directorate is already funding five other synthesis centers.

In sum, these concerns point to the fact that we need to find a short and effective path to getting glowing reviews on basic data availability.

As NSF thinks about how to help us address these concerns, Dr. Twombly sought to gather specific information from Information Managers, the tasks we face, and the impediments we have to doing our jobs on data availability. She asked about our job responsibilities, our setting of priorities with respect to the processing of data sets, our interactions with LNO, other network Information Managers, site PIs, and about our strategies for balancing site and network demands. She also asked us about our expectations for the NIS (by which she was referring primarily to PASTA). She also asked about the specific impacts on us by the Macrosystems Proposal data management plans and the new overall data management plan requirements. The overall concern was to understand how we manage our workloads and conflicting goals to maximize data availability.

IM Exec responded first by describing our overall job responsibilities. We pointed out that, while responsibilities vary across sites, we are all faced with assuring the availability and reliability of a basic stack of Information Technology architecture. We noted that while not all of us are directly responsible for the cabling in the walls, for example, some of us are directly responsible for such things and all of us have to give some thought to the fact that those cables are and will continue to be available. The same is true for the areas of system administration, database administration and design, web site administration and design, etc. In addition, we pointed out that many of us are actively involved in producing some of the long-term data sets that are placed on line and that many of us are involved on the executive bodies of the sites we serve, providing the perspective needed to make sure IT infrastructure meets research needs.

We also noted that while we do have apparently conflicting demands in terms of our site and network responsibilities, the significant amount of network interaction among the Information Managers reduces our workload over time by both providing a support group from whom we can learn and borrow tools and approaches and by providing a community standard against which we can design our systems. Without these systems, every IMS at every site would have to be designed from scratch, resulting in significant more expense in terms of both time and money. As examples, we pointed to our recent collaborative use of IM supplement funds for the development of projects like the Unit Registry, the Controlled Vocabulary, DEIMs, cross-site of adoption of the GCE Metabase, and, most notably of late, the SensorNIS efforts, which have resulted in unprecedented cooperation and sharing of efforts within and even beyond the LTER. We also pointed out the fact that although PASTA is a component of the NIS, the Network Information System itself is a much larger, modular, and distributed product that includes SiteDB, PersonnelDB, ClimDB/HydroDB, the LTER All-Site Bibliography, etc. Many of the components of the NIS were and continue to be developed and supported by site-level Information Managers.

In response to our emphasis on the supplement-funded projects and the importance of IM-specific supplements, Dr. Twombly informed us that the NSF is currently thinking about different ways to use the supplement money and wondering whether or not it might be more useful to set aside supplement funds in a way that would allow them to support larger, proposal-driven IM projects through bids collected in the fall, rather than having smaller amounts be offered directly to the sites during the later summer. This is a question we promised to take back to both the IMC and our Site PIs. However, we did note that a current challenge facing the IMC is the fact that

while many excellent tools do "bubble up" from site IMs for use across the network, we seldom have the funds to bring those projects into full network production and then to follow up and support their adoption across sites.

Saran also took the opportunity to mention to us, that during the site review for LNO, scheduled for mid-May, she will be arranging for reviewers to do phone interviews with representatives from several LTER sites, including IMs and PIs. Site IMs may be contacted as part of this process.

Meeting with NSF Program Officers, Dr. Peter McCartney and Dr. Saran Twombly (attended by IM Exec members and Kristen Vanderbilt)

Dr. McCartney echoed and re-emphasized many of the points brought up by Dr. Twombly. He noted that the LTER Network needs to find a way to get credit for something that it often does well, which is general data management. The problem at the present time is that by emphasizing PASTA, we are setting ourselves up to reviewed against something that, rather than being a practical short-term outcome, is instead an overall ultimate goal. Such expectations are unrealistically high and we are bound to fall short of them. Due to the fact that provenance awareness and related technologies are advanced tools that are still in their research stages, we need to re-focus our energies on basic data availability.

The most common complaint we face in reviews is that our data catalogs are not complete. We do not need to build for a situation in which people who have no idea what they are looking for should be able to find useful data. According to Dr. McCartney's experiences, 99% of reviewer scenarios involve a scientist reading a publication and then turning from that publication to either a site or centralized catalog in order to find the data associated with that publication. This is quite different from someone who is doing exploratory research and this is the scenario for which we should be building. In this sense, the test from the Robbins video is not a fair one – and this fact is a point Dr. McCartney continues to make to colleagues in the NSF. However, it does raise general issues related to the relative merits of site-level vs. centralized data and metadata repositories.

At this point, Dr. McCartney feels that we need to focus on two questions:
1. What we can build with the resources we have?
2. What can we build that will do something really great with regard to data availability?

Moving forward, NSF would like to see LNO and IMC on the same page with regard to the solutions we pursue. The IMC needs to work with LNO to set priorities. The IM Exec responded with overall agreement on these perspectives. We did point out, however, that some short-term gains in data availability without building complete metadata may come with long-term costs in terms of data value, due to the lack of context for future analysis. The further one gets from a data set's originator in time, space, or social distance, the less likely it will be that you can understand the necessary context, and a balance is necessary. With Dr. McCartney, IM Exec again pointed to some of our shared solutions (such as DEIMS and the Metabase) as possible avenues for improvement. We also noted that the LNO Metacat Instance has improved a great deal in a short time and we suggested that one avenue to help PASTA succeed would be to prioritize key data sets that could be ingested into PASTA. The goal would be to focus on data that would, because of the nature of the data they contain, demonstrate PASTA's value.

The topic of older data was again raised and we briefly discussed the Legacy Data Project. At present the proposal resides with NSF and we await further information from LNO or the EB.

White Paper

Given the discussions in the meeting, it was determined that the white paper discussed during our 30-year review response should relate the overall IMC and LTER vision for data availability. In terms of the structure of the paper and the work plan, IM Exec agreed that white paper topics should first be presented in a session at ASM and that those presentations, combined with feedback, should be distilled into the final product. The audience for the white paper is to be the entire LTER community and NSF and it should be grounded in and understanding of broader culture and paradigm shifts with relation to the use of information in western society in order to establish context for LTER's needs. John Chamblee will take the lead in the outline, with help from Margaret O'Brien, Don Henshaw, Emery Boose, Wade Sheldon, and Corinna Gries. The outline and a call for authors should go out in May, the workshop presentations should be assembled by August and the first draft should be done by the end of October.

Summary Discussion

After our meetings with Scott Collins and Bob Waide, as well as with Drs. Twombly and McCartney, there was consensus among all members of IM Exec, that the information we've received from the NSF demonstrates the need for an immediate and dramatic directional change with regard to the way that we are addressing network-wide challenges to data availability. In the past several years, we have been focusing on a multitude of parallel projects that are all moved forward at the same relative pace. Each project is designed to increase not only data availability, but also data quality and the overall interconnectedness of the network and the information systems supporting it. While we know that these are valid long-term goals, we also recognize that, in the short-run, we must emphasize data availability and set priorities that will allow rapid progress in this area. This means that other goals will have to be temporarily de-emphasized in order to focus concentrated energy on data availability. IM Exec has drafted a preliminary plan to identify the approaches that will yield quantifiable results with regard to data availability. We believe that, once modified by feedback from the IMC, NISAC, the EB, this basic outline will help the LTER Network move forward rapidly toward the common goal of shifting LTER's emphasis back to scientific pursuits.

The plan is presented as an outline of activities that are presented in order of priority and are provided with the following sets of additional explanatory information:

Context: the feedback we received, or the current state of affairs.
Goal: a possible solution to perceived problems given the context information.
Strategies: how IMexec thinks we could address the goal (tasks tbd).
Funds: funding goal that could support the tasks (key below table). SIP id: reference to Strategic & Implementation Plan.
Who/Where: responsible parties for accomplishing the tasks.
Related issues: other comments or issues which are likely to come up.

Above the table is a list of fundable projects that would help move these goals forward. Below the table is a list that defines what Scott Collins and Bob Waide believe to be the "key features of data packages" that, if present, will improve data accessibility. IM Exec wants to work LNO and the EB to refine these features and make plans for assuring their existence at sites.

Fundable projects:
1. EML SWAT Team/mentors
2. DEIMS Developer
3. Metabase Developer
4. Data catalog improvements

	A		B	C	D
Context	We have data with metadata (ie, in Metacat) where data are there, but are still not obvious		There are data that are not yet described	People expect to find some kinds of data from all or most sites (eg, CRA)	NSF: PASTA data could be prioritized
Goal	More existing data packages have visible URLs		More data described	Derived data that addresses science priorities is emphasized	Data intended for PASTA are identified.
Strategies	1. Adapt existing EML 2.Identify type II (ie, no URL) 3.ensure the 5 essential dataset features (below)	Further improvements to current data portal presentation might be necessary	Prioritize backlog Data expedition	Identify high priority derived data Encourage sites to contribute PASTA-ready input data Highlight those products and workflows.	Possibilities include: 1. data packages needing workflows 2. time-series
Funds	1, 2, 3	4	1,2	4	
SIP id	1A, 2BC		1C (all), 1E (all)	1AB	1AB, 1B (all)

Where, who	At sites. Ideally, new temp, or EML-SWAT	At LNO, Probably a temp	Site IMs and scientists	NISAC, Sites, LNO	NISAC
Related issues, notes	The use of data registration forms will come up, (sip 2BB)		This is a long-term activity.	Informs column D priorities	How is the NIS planning to handle non-PASTA data?

Essential features of LTER EML data packages to improve discoverability and access
The first three features are used for full-text searches. Titles and abstracts should also be designed for human readability.

1. **Titles** - The dataset title should be descriptive, mention the data collected, geographic context and research site (what, where), and possibly, the time frame (when).

2. **Abstract** – Include an abstract rich with descriptive text, analogous to a paper's abstract. Taxonomic information may be appropriate. This is a good place to indicate whether the dataset is ongoing or complete. Some general terms regarding methods, instrumentation or measurements should also be included.

3. **Keywords** - Since keywords are searched in LTER queries include a meaningful set of keywords identifying the LTER site and research context, a set of keywords from the LTER controlled vocabulary (http://vocab.lternet.edu), and a set for the LTER core research areas. This is also a useful place to add additional terms which do not fit into data package titles or abstracts.

4. **Data Table Description** - Data table descriptions for tabular data should be complete, including attributes and physical format with a data distribution URL.

5. **Data distribution URL** - This URL to the data is located with the data table description above. The URL should deliver a data stream and not point to another application or use page. Web views reflecting LTER data availability are optimized for a download URL at this location.

If sites don't already have data table information in EML or some other structured form, items 4 and 5 are nontrivial, and in a different class of effort from the 3 above. All five of the features listed are taken directly from the EML Best Practices for LTER Sites Version 2 and assume the site is able to produce well-formed EML that is compliant with the XML schema, the EML schema, and the EML ID and reference structure. The majority of LTER data are tabular, and tables are assumed here. However, these features can be applied to other data types such as spatial data (e.g., GIS).

Action Plans

1. IM Exec will produce a detailed account of this meeting and post it on-line in order to discuss it with the rest of the IMC as soon as possible. This may involve rescheduling VTCs and will certainly involve discussions of the above plans as well as the need to seek feedback regarding

the use of supplement funds.

2. IM Exec work with Bob Waide to clarify and help develop plans for use of the $170,000-$200,000 dollars available to fund projects directed at increasing data availability.

3. The IMC Co-Chairs will compile reports for NISAC and the EB.

4. IM Exec will compile a letter to Scott Collins regarding Co-Chair compensation funds.

5. IM Exec will prepare the IMC annual meeting agenda.

6. John Chamblee will begin outlines on the White Paper and a related ASM workshop.

Meeting Notes [2]

Source URL: http://im.lternet.edu/news/committees/im_exec/notes/2012_04_3

Links:
[1] http://im.lternet.edu/user/27
[2] http://im.lternet.edu/taxonomy/term/3