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

Home > IM Exec > IM Exec Winter meeting 2010-02-23/24 (Albuquerque, NM)

IM Exec Winter meeting 2010-02-23/24 (Albuquerque, NM)

Thu, 03/18/2010 - 12:55pm — administrator [1]

Tuesday Feb 23 0800 - 1700, Wed Feb 24 0800-1700 @ UNM Albuquerque (Ceria)

Participants
IM-Exec: Don Henshaw, Margaret O'Brien (co-chairs), Sven Bohm (VTC), Emery Boose, Hap Garritt, Corinna Gries, Suzanne Remillard
IM guests (occasional): Karen Baker, Wade Sheldon (VTC), Kristin Vanderbilt
LNO staff (occasional): James Brunt, Duane Costa (VTC), James Moss, Mark Servilla, Bob Waide, Yang Xia
Other guests: Todd Crowl (VTC), Christine Laney, Peter McCartney (VTC), Deanna Pennington

Topics
1. Governance
2. WG Planning
3. CI-Team Project
4. LNO Report
5. 2010 IMC Meeting
6. Workshop Schedule
7. Working Groups
8. NISAC Report
9. NSF VTC
10. EML Metrics
11. Upcoming Events
12. Action Items

1. Governance

The role of IM-Exec needs to be clarified. To date IM-Exec has been a steering committee, but future NIS development will require regular input and decisions from the IMC. The IMC could formally authorize IM-Exec to represent it and to make decisions on its behalf, as the EB now does for the SC. Other details related to IM-Exec (membership, terms, etc) also require clarification. These issues will be considered by the governance WG and discussed at the IMC meeting next fall. One solution would be to create bylaws for the IMC. The original LTER governance committee might also be reconvened to consider the role of the IMC, IM-Exec, NISAC, and the EB in the new LTER governance structure.

2. WG Planning

NISAC has proposed that "tiger teams" be created to work closely with NIS developers at LNO. The tiger teams would be expected to interact with LNO developers in a timely manner. There may be funding to compensate tiger team members. The LNO operational plan includes funding for two training workshops and two production workshops each year. Production workshops will be expected to produce a product. We are well into the first year so a schedule for these workshops is needed. If desired, the workshops could be front-loaded.

Science WGs should include at least one IM. Phil has asked James to participate in the four prospectus WGs. Lynx and Mirada are examples of cross-site projects without extensive IM planning. One possibility would be to add an IM component to the RFP process.

3. CI-Team Project

The CI-Team program at NSF is funded by the office of cyberinfrastructure. CI-Diffusion is a new program that addresses the problem of communicating across boundaries (scientists, information specialists, etc). The LTER network is an ideal application. Each institution can submit only one proposal. Deanna is preparing a prospectus for the internal competition at UNM (due Friday). The full proposal is due at NSF on April 25. If funded, Deanna's work would focus on how to bridge perspectives and find common ground. Training would be conducted mostly via VTCs with some training during scheduled meetings. Targets would include the NIS, NEON, and DataOne. There have been positive responses from some PIs and letters of support would be helpful.

4. LNO Report

The LNO operational plan was written to encourage broad participation and refers to the IMC. IM-Exec is probably the best mechanism for providing IM input, but there also needs to be way to incorporate minority views. Bob would like to see a schedule for submission and review of all proposals, ideally once a year in the fall.

Existing pots of money include (per year): $100k for research WGs, $50k for mini-sabbaticals, $170k for decadal plan WGs (including prospectus WGs), $20k for production workshops, $43 for training workshops, and $15k for IM release time. There were no separate funds for post-ASM workshops this time. The ARRA award to LNO provides 6 years of funding distributed over 5 years (essentially a 20% contingency).

There is an ongoing effort at NSF to get more funding for individual sites. In January an ad hoc LTER committee led by Peter Groffman (including Don, Emery, James & Wade) completed a prospectus to NSF for possible use of NEON funds to address LTER legacy data issues. There has been preliminary support from NEON. The proposed budget includes a total of $200-600k per site over a period of 3-5 years. NEON has suggested adding value-added datasets.

The unit dictionary proposal has been funded as a post-ASM workshop. The Drupal training proposal is waiting for a recommendation from IM-Exec. Two RFPs have come out at NSF but may not be a good match for the prospectus WGs. Proposals that do move forward will need to engage IMs. No immediate tiger team needs have been identified. Jonathan has volunteered to be IM rep for the communication WG.

The next SC meeting is scheduled for May at PIE. IM-Exec may be invited to attend. Prospectus WGs and the NIS are on the agenda for the morning of the first day with a field trip in the afternoon. The second day will be devoted to creating an LTER strategic plan, which will include a diversity plan, what the network should look like (number of sites and where), and a communication plan for LTER (including a plan for the NIS). A facilitator may be hired. The next

SC meeting in Moorea (2011) will be too expensive to add other participants.

5. 2010 IMC Meeting

The next IMC meeting is tentatively scheduled for 21-23 Sep at KBS. The three main topics for discussion will be: governance, EML metrics and best practices, and network database redesign and web services. The tentative schedule is:

Tuesday evening: mixer & rotating governance discussion.
Wednesday: 8:30-10:30 governance, 11-1 EML metrics, 1-2 lunch, 2-4 database redesign, 4-6 field trip, 6-7 dinner, 7-9 demos.
Thursday: 8:30-10 reports, 10-12 small groups (GIS, controlled vocabulary, others as needed), 12-1 lunch, 1-3 business meeting & NSF VTC, 3-5 reports & wrap-up, mixer, dinner.
Friday: 8-10 IM-Exec meeting.

WG membership lists could be kept up to date on the IM website. IM-Exec needs to set an example for use of the website.

6. Workshop Schedule

Workshops (TT = tiger team, PW = production workshop, ASM = post-ASM workshop):
2010: TT data suite phase 1 (EML metrics?), TT discovery / access API, TT data access portal, PW EML best practices, PW network redesign, ASM units DB completion, ASM LTER maps.
2011: PW ClimDB integration.

IMC Annual Meetings:
2010: governance (Karen, Nicole, Eda), EML best practices (Margaret), network DB redesign (Corinna, Mason), controlled vocabulary, GIS.
2011: EIMC meeting

IM Training:
2010: tiger team preparation, Drupal installation.
2011: EML metrics / data portal, web services for consumers (units=example).

7. Working Groups

Active WGs include: unit dictionary, controlled vocabulary, governance, GIS, EML metrics, and web services. Targeted WGs include EML best practices, projectDB, and IM website. Drupal forums has not been used to date but could be linked to email lists. ProjectDB may become a NIS module.

8. NISAC Report

NISAC has been engaged in development of the LNO operational plan and setting milestones for a year-end review of the LNO (as tasked by the EB). NISAC will be asked to review progress on the NIS. NISAC is also planning the agenda for the upcoming March meeting. Topics will include scope and governance issues as well as formalizing recommendations for the CIIP and LNO (including tiger teams). In addition NISAC would like to come to a meeting of minds with IM-Exec regarding details of the NIS implementation plan.

There is a growing awareness of our need to approach standardization in many areas. The EB is looking to NISAC and IM-Exec for how to proceed (esp. actionable items). NSF and the EB may have a false expectation that standardization can be applied at the IM level. Ultimately this is an

issue for the scientists. There is EB support for a tiger team on standardization. At present LTER has no enforcement mechanism. New collaborative projects present the best opportunity for standardization. Standardizing legacy data is difficult at best.

We need a framework in which to align activities. WGs can supply a context for tiger teams to work out details. Earlier CIIP and decadal plan work was largely ivory tower; now with ARRA funding we need to work out implementation details. NISAC is expected to identify tiger team needs with input from IM-Exec and the EB. Finding enough capable individuals in LTER to staff the tiger teams may be a challenge. Compensation for tiger team leaders, IM release time, etc are under discussion.

NISAC foundational documents need to be updated. Recent requests from the EB have extended the original scope of the committee. Clarity is also needed on membership and terms. Engaging science members of the committee has been a challenge. The scientists leading the prospectus WGs might be good candidates. Many lead PIs still feel that technological developments at LNO, NISAC, etc are too independent of research science.

The governance structure of NISAC is open to discussion. In the past the IMC has recommended (and the EB has confirmed) scientists for NISAC membership. The role of the domain scientist co-chair has not been well defined. In general this position has not worked out as expected as a liaison to the SC. Wade will step down as IM co-chair this spring. IM-Exec welcomes recommendations for his replacement.

9. NSF VTC

No word yet on supplements. Todd is still waiting for a 2010 LTER budget. Data and schoolyard remain the highest priorities. The LTER legacy data prospectus is an exciting development. A strategic plan could help position LTER for larger amounts of funding.

NSF is looking for a synergy between NEON and LTER. Peter has encouraged NEON to share its CI plans though there are some constraints because of the competitive bidding process. NEON needs to build according to a cost schedule (unlike research projects) so it doesn't have much room for R & D. Some in LTER are looking to NEON to solve sensor network problems. But NEON will use (very expensive) off-the-shelf software wherever possible. In-house open-source solutions are not always cost effective for large systems. Todd and Peter recommend a meeting with Tony Beasley to see what NEON is up to.

NEON datasets are larger but simpler than LTER datasets. NEON has accepted rigorous limits on its data. This is a compromise that LTER, NCEAS, etc have not been willing to accept. But it avoids many difficult problems (semantic mediation, etc). LTER might consider adopting a more NEON-like approach. There are few successful examples of data integration in LTER. The two most recognizable accomplishments are ClimDB and EcoTrends. The new synthetic initiatives will be an opportunity to add to these.

RFPs for the new macro-scale ecology initiatives will ask for a detailed data management plan (1-3 pages). Data management may eventually become a third criterion (along with intellectual merit and broader impacts).

Of 60 NEON tower sites, at least 16 are LTER sites. This overlap provides a great opportunity to work out legacy data issues. Maintain both lines of effort: value-added products and access to all original data.

NSF ABI (advances in biological informatics) is looking to support R&D proposals with a clear

biology focus. The program would like to see more proposals from ecology on what we can do with the data (and not just how to manage it).

NSF is writing its own strategic plan for LTER. LTER is a proven concept. There is some trepidation about NEON at NSF.

All site reviews last summer noted great strides in data accessibility. During the reviews Todd searched online for datasets to match the talks. The PIs and scientists have gotten the message. Twelve renewal proposals have been submitted for a panel in April.

NSF is working toward a foundation-wide policy for data management. NSF is seeking to stimulate a peer-review process, rather than a top-down policy like NIH. The best practices and review documents in LTER have been very helpful.

EML is a community standard without explicit funding. How to maintain such initiatives? Submitting successive proposals is not a sustainable solution. Not all software can or will be commercialized. Peter Arzberger (new DBI head) will try to address this issue.

10. EML Metrics

Current checks on EML are XML schema compliance and the EML parser (ids and references). EML comes with a Data Manager Library (in Java) which can read data entities into a RDBMS. The DML provides additional checks on EML but has limited reporting capability. If reporting were improved, the DML could be a valuable resource for checking EML and associated data.

Potential problems include: parsing EML, returning a data entity, parsing the data entity, ingesting it into a RDBMS. Problems may be easy to fix or may require improvements to the site information management system. The goal is create a list of criteria, initiate software tools, and generate initial reports on site data (if possible by the Sep IMC meeting).

The DML was initiated in Aug 2006 by Jing Tao and Duane Costa. Ben Leinfelder did some upgrades for another project. The Semtools project will use it but does not plan to upgrade. Future development will be accomplished by LNO programmers under Mark's supervision. In the PASTA diagram, the DML is part of the Data Loader. The DML will be the point of contact to register datasets for loading and will provide a portal for quality checks, etc. Currently the DML handles only data tables and will need to be extended for other data types. These details will need to be fleshed out by a tiger team.

The DML was originally designed to cover four use cases (implemented) and two extended use cases (not implemented): (1) Parse metadata (EML or other). This need arose from Kepler. (2) Download data from a remote source to a local data store. (3) Load data into a relational database table. Currently supports HSQL, Oracle, and Postgres. (4) Query data from the relational database. Queries can be stored in a property file. (5) Set upper limit on size of database. (6) Set lifespan on individual data tables. The DML was originally designed as a black box that just does its thing. A report function (aside from occasional error codes) was not anticipated.

Extension and hardening of existing code by September is probably doable. But not the final product that will be part of PASTA.

See Christine's power point for salient problems encountered by EcoTrends. EML often lacks key information about how the data were collected and processed (e.g. preprocessing of sensor data). The DML could indicate the presence or absence of a method tree.

The interface should be a web page like the current EML ID and references parser (http://knb.ecoinformatics.org/emlparser/). One interface and two steps (metadata and data). Structural (but not semantic) checks on data will be possible.

The system could check on EML completeness according to best practices. A detailed list of criteria would be required. This would go beyond schema compliance and would be LTER specific. Missing fields could also be identified using a style sheet. Note that EML 2.1 (unlike EML 2.0) does not permit empty strings.

The tool should accept a list of elements as a parameterized list. Check that the data URL is valid and (if function=download) returns a data object. PASTA is intended to support persistent identifiers (e.g. DOI) and will follow community best practice.

Possible items to check: number of header lines, number of rows & cols, delimiters, missing value codes, reserved terms, duplicate variable names, data typing, number typing, duplicate records, URL validity, character encoding, controlled vocabulary. Expose downloaded data to end user as a sanity check. Ideally the tool should go through the entire table and report all problems. At present the DML aborts if a problem is found that would preclude successful data harvesting.

The LNO operational plan indicates that NISAC and the IMC will help define tiger teams. Tiger teams will be in regular contact with LNO developers. Iterative development cycles may involve tiger team review. A more detailed management plan is needed. Build on existing collaborations and reuse technology where possible.

The EML metrics WG will follow up with a VTC to organize these ideas and reconnect with Mark et al. later in March. After the mini-review, LNO will flesh out milestones for the next couple of years. According to the LNO operational plan, the schedule for production workshops will be determined by the IMC, NISAC, and the LNO. Training proposals go to the EB because they may involve scientists.

SKOS (simple knowledge organization system) = http://www.w3.org/2004/02/skos/.
HIVE (helping interdisciplinary vocabulary engineering = http://ils.unc.edu/mrc/hive/.

11. Upcoming Events

SC meeting. May 12-13, 2010 at PIE. Travel and EB meeting on May 11.

IMC meeting. Sep 22-23, 2010 at KBS. Travel on Sep 21.

EIMC meeting. Fall 2011 in Santa Barbara.

Water coolers for 2010. Supplements = ASAP. Mar 1-2 = IM-Exec update. Apr 5-6 = Units. May 3-4 = inter-site IM projects. June 7-8 = governance. July 6-7 = NISAC update. Jul 19-20 = EML metrics. Aug 2-3 = LNO update. Aug 16-17= NIS update. Sep 7-8 = network database redesign.

12. Action Items

Write annual IMC report to EB
Complete IM-Exec meeting notes
Respond to Drupal training workshop request
Invite John P. to discuss controlled vocabulary
Recommend additional IM for communication WG
Follow up on IM-Exec attendance at SC meeting

Create PDF of 2009 IMC meeting report
Update IMC WGs on website

Related Links

- Deana Pennington's current project with NSF/CI-TEAM (443) [2]
- EML Specification (270) [3]
- LNO suggested training topics (213) [4]
- NISAC CIIP presentation at ASM 2009 (292) [5]
- use cases for EML data manager (code) library (490) [6]

Attachment	Size
Logistics_imexec_20100223.doc [7]	43 KB
EML_210_Specification.pdf [8]	1.67 MB
WG proposal, Sept 2009: EML_metrics_wg.doc [9]	32 KB
Christine Laney's 2009 presentation featuring the EcoTrends experience with LTER data: EcoTrendsPastPresentFuture.pdf [10]	3.16 MB
Stawman planning spreadsheet [11]	18.5 KB
Introduction to EML metrics session (EML_metrics_intro.ppt) [12]	113.5 KB
ChemDB project - Feb 2010.ppt [13]	605.5 KB

Meeting Notes [14]

Source URL: http://im.lternet.edu/news/committees/im_exec/notes/2010_02_23

Links:
[1] http://im.lternet.edu/user/1
[2] http://im.lternet.edu/links/goto/558/99/links_related
[3] http://im.lternet.edu/links/goto/558/64/links_related
[4] http://im.lternet.edu/links/goto/558/101/links_related
[5] http://im.lternet.edu/links/goto/558/100/links_related
[6] http://im.lternet.edu/links/goto/558/102/links_related
[7] http://im.lternet.edu/sites/im.lternet.edu/files/Logistics_imexec_20100223.doc
[8] http://im.lternet.edu/sites/im.lternet.edu/files/EML_210_Specification_0.pdf
[9] http://im.lternet.edu/sites/im.lternet.edu/files/EML_metrics_wg.doc
[10] http://im.lternet.edu/sites/im.lternet.edu/files/Laney_EcoTrendsPastPresentFuture.pdf
[11] http://im.lternet.edu/sites/im.lternet.edu/files/IMC_5year_plan_brainstorm_4.xls
[12] http://im.lternet.edu/sites/im.lternet.edu/files/EML_metrics_intro_0.ppt
[13] http://im.lternet.edu/sites/im.lternet.edu/files/ChemDB project - Feb 2010.ppt
[14] http://im.lternet.edu/taxonomy/term/3