LTER Information Managers Committee Report

ASM 2009, Estes Park, CO

13 Sept 2009

1. Participation:

1.1 <u>US LTER, by site</u>: Margaret O'Brien (SBC), M. ("Gastil") Gastil-Buhl (MCR), Dan Bahauddin (CDR), Wade Sheldon (GCE), Hap Garritt (PIE), Christine Laney (JRN), Donald Henshaw (AND), Suzanne Remillard (AND), Mason Kortz (PAL/CCE), James Conners (CCE/PAL), John Campbell (HBR), Jason Downing (BNZ), Jamie Hollingsworth (BNZ), Linda Powell (FCE), Adam Skibbe (KNZ), Theresa Valentine (AND), Corinna Gries (CAP), Sue Welch (MCM), Karen Baker (PAL/CCE), Emery Boose (HFR), Dave Balsiger (NTL), Lynn Yarmey (CCE/PAL), Jim Laundre (ARC), Barbara Benson (NTL), Steve Jackson (KBS), John Chamblee (CWT), John Porter (VCR), Eda Melendez (LUQ), Florence Millerand (PAL/CCE), Ken Ramsey (JRN), Kristen Vanderbilt (SEV), Sven Bohm (KBS), Hope Humphries (NWT), Barbara Nolen (JRN), Jonathan Walsh (BES), John Campbell (HBR)

1.2 <u>LTER Network Office</u>: John Vande Castle, Inigo San Gil (and NBII), Mark Servilla, James Brunt, Duane Costa

1.3 <u>International LTER by site or country</u>: Akiko Ogawa (JAPAN LTER), Helena Karasti (FIN LTSER), MeeiRu Jeng (TERN Taiwan), David Blankman (Israel LTER), Victoria Goodall (SAEON), Sabina Datcu (Romanian LTER), Chau Chin Lin (TERN Taiwan)

1.4 <u>Guests</u>: David Ribes, Jim Gosz (University of Idaho), Mark Schildhauer (NCEAS), Derik Barseghian (NCEAS), Ben Leinfelder (NCEAS), Bruce Wilson (ORNL), Todd Crowl (NSF), Henry Goltz (NSF)

2. Business meeting:

2.1 <u>Outgoing IMC Co-chairs</u> and organizers of this meeting: Corinna Gries (CAP) and Nicole Kaplan (SGS)

2.2 <u>IMC Co-chairs elected (term ends 2012)</u>: Margaret O'Brien (SBC) and Don Henshaw (AND)

2.3 <u>IMExec membership</u>: Sven Bohm (KBS 2012), Emery Boose (HFR 2012), Corinna Gries (CAP 2011), Suzanne Remillard (AND 2011), and Hap Garritt (PIE 2010), James Brunt (LNO, ex officio), Wade Sheldon (GCE, intermittent as NISAC co-chair). A policy was agreed upon by the IMC that IMExec will maintain 7 members. When the chair position is shared, there will be a year in which only one IMExec position is open instead of the usual two.

2.4 <u>Executive Board liaison:</u> Corinna Gries (CAP)

2.5 <u>NISAC membership (IMC maintains 4 members)</u>: Wade Sheldon (GCE, co-chair), Karen Baker (PAL), Kristin Vanderbilt (SEV), and John Porter (VCR).

2.6 <u>Databits editor(s)</u>: Jason Downing (BNZ 2010), to be followed by Dan Bauhauddin (CDR 2011).

2.7   New Information managers: Corinna Gries moved from CAP to NTL, Sue Welch (MCM), John Chamblee (CWT), Hope Humphries (NWT).

2.8   Open IM position: CAP

3. NISAC report and discussion on CI Strategic Planning

3.1   Organizer: Wade Sheldon

3.2   Presentation available at:
      http://intranet.lternet.edu/im/files/im/NISAC%20CIIP%20IMC%20Working%20Group%20ASM2009

3.3   Executive Summary: Wade Sheldon (NISAC co-Chair) presented an overview of the Cyberinfrastructure Implementation Plan (CIIP) currently under development by the Network Information System Advisory Committee (NISAC). The draft describes near-term, mid-term and long-term tasks planned for each CI initiative described in the Decadal Plan. The CIIP will provide specific tasks for the broader vision of the Decadal Plan, but will need to become a living framework that is revised over time based on experience and redirection of vision.

4. IMC Working Group activities (current):

4.1   **Governance** – discussion by all IMC members and some guests

4.1.1   Organizers: Nicole Kaplan and Karen Baker

4.1.2   Full report available at:
        http://intranet.lternet.edu/im/news/meetings/2009/workgroups/governance

4.1.3   Executive summary: Discussion focused on the decision-making processes and communication among the committees and boards across the LTER organization, and highlighted the partnership between IMC and LNO when the LNO staff is an active participant or developer on a working group. It was recognized that for recent projects, we have moved from informal partnerships to a semi-formal consensus method, and may now need additional formal steps, particularly at the point when resources are assigned. NISAC has developed requirements for potential NIS modules (2005, http://intranet.lternet.edu/modules.php?name=UpDownload&req=getit&lid=94), and this document is likely to need update and revision. Science working group projects go through an approval process with the EB, while LNO manages funds for the IMC and IM projects, although it is unclear how projects are vetted. Since 2005, network governance has changed so that EB assigns resources - an arrangement which means that technical proposals may not be vetted through a group with the appropriate expertise. A revision to this procedure may be required so that decision-makers are informed about the scope, ramifications, and coordination of the project in the network framework. There was wide agreement that all projects (NIS or otherwise) should address design, prototyping, testing, maintenance, update, site-level enactment, training, and other topics as needed. Additionally, projects should be defined according to specific criteria and to their relationship to LTER and NIS components such as "LTER-affiliated", "LTER-endorsed", "NIS core, in production", "NIS core, in development", etc.

4.2   **GIS Data** – break out

4.2.1   Organizer: Theresa Valentine

4.2.2   Full report available at:
        http://intranet.lternet.edu/im/news/meetings/2009/workgroups/gis

4.2.3   Executive Summary: The group has composed a draft document "Proposed GIS
        Recommendations for LTER Sites" to be reviewed and revised. The next likely
        step was surmised to be a formal presentation of the recommendations to IM Exec
        and/or NISAC, which is likely to depend on resolution of communication steps
        discussed by the governance group. The GIS group recognized that not all site
        possess the expertise required to meet potential proposed recommendations and so
        GIS and Remote Sensing Training will be required. This group plans to:
        a. outline potential training opportunities, including (but are not limited to): data
        discovery, server and internet mapping applications, GIS certification, modeling,
        emerging API's, data integration (e.g., between GIS, remote sensing, LiDAR),
        Google Maps (particularly with LTERmaps)
        b. conduct a survey of web mapping tools and skills at LTER sites, and to
        c. propose a workshop during the winter of 2010. Please note other GIS
        opportunities that were available at the ASM, e.g., Workshops and the GIS Lunch.

4.3   **Controlled Vocabulary** – break out

4.3.1   Organizer: John Porter

4.3.2   Full report available at:
        http://intranet.lternet.edu/im/news/committees/working_groups/controlled_vocabulary

4.3.3   Executive Summary: To date, this group has assembled a list of selected terms
        from LTER EML documents, which has been cross-linked to major keyword
        dictionaries, edited to the preferred form with place names and taxa removed, and
        reviewed by IMC via web survey. The result is 640 words (148 synonyms, 201
        NBII keywords, and 21 GCMD keywords). The next steps are not entirely clear,
        and like the other groups needing to advance projects, depend on the model which
        best suits the IMC and network CI needs. This project is analogous in scope and
        use to the unit dictionary, and the two should be advanced along similar paths.
        Tools will be necessary. An AJAX auto-complete Metacat interface was completed
        in 2009, and a prototype keyword-addition tool for EML was demonstrated.
        Discussion topics included:
        a) the relationship of LTER terms to other keyword and vocabulary activities (most
        of which are more narrowly focused than LTER's), including: SONET is mapping
        between ontologies and observational models (Gries). Semtools is adding
        ontological annotation to EML with Morpho (O'Brien). NASA has developed
        SWEET (Semantic Web for Earth and Environmental Terminology).
        b) Are there LTER EML documents which is not attached to any keyword? Find
        these, and consider site-specific keywords.
        c) The need for other methods of handling taxonomy and place names, topics
        which also have relationships to other WG at this meeting.

4.4   **Units Dictionary** – break out

4.4.1   Organizers: Lynne Yarmey and Mason Kortz

4.4.2   Full report available at:
http://intranet.lternet.edu/im/news/meetings/2009/workgroups/unit_dictionary

4.4.3   Executive Summary: Two major efforts were defined as the early focus of this group and are largely completed: the Unit Dictionary itself (comprising the set of units in use by the network and the best practices that cover them), and the Unit Registry (the application used to access and manage these units). In the dictionary, a unit may be scoped at different levels, a model which facilitates additions that are not yet vetted by a unit-authority or against best-practices. The Registry prototype at PAL/CCE (UCSD) was demonstrated, showing the web service to retrieve units, quantities, and scopes, plus its unit-validation service. Discussion focused questions about the current service and with suggestions for new features and possible site- and network-level use. The most immediate activities of the group will be:
a) to review current contents of the Dictionary (EML and site) for compliance with the draft best practices,
b) to review and revise the best practices document itself. Volunteers were also recruited to implement the dictionary and registry at their sites, providing feedback and bug fixes.

5.   Working Groups (new):

5.1   **Web services** – newly established IMC working group

5.1.1   Proposed members: Mason Kortz (lead, CCE/PAL), Suzanne Remillard (AND), Wade Sheldon (GCE), Jonathan Walsh (BES), John Porter (VCR), Sven Bohm (KBS), Dan Bahauddin (CDR), Jason Downing (BNZ), Adam Skibbe (KNZ), Corinna Gries (NTL), James Conners (PAL/CCE), M. Gastil-Buhl (MCR), Mark Servilla (LNO), James Brunt (LNO)

5.1.2   Executive Summary: The IMC would like to make use of databases at the LNO, preferably via web services. A working group was established with developing recommendations for web services and integrated databases at the LNO. Possibly this will entail multiple groups or timelines; for example, a second group could be assembled to outline management and curation strategies for integrated network databases in light of the fact that parts of the software framework are in need of updating. Concrete tasks are in development, and might include:
a) Develop guidelines for creating and documenting network-level web services and the LTER SOA,
b) Develop recommendations for a common service interface to existing LNO databases, and to which new databases could be integrated,
c) Identify needs in the community that could best be addressed through network web services, and report to IMExec/NISAC on these needs.
d) Identify site projects that could evolve into network-level web services,
e) Help sites implement web services, either as consumers or providers.

5.2   **EML Metrics** – ad hoc ASM work group

5.2.1   <u>Proposed Members</u>: Margaret O'Brien (lead, SBC), Emery Boose (HFR), Dan Bahauddin (CDR), Jonathan Walsh (BES), James Brunt, (LNO), Mark Servilla (LNO), Duane Costa (LNO), Mark Shildhauer (NCEAS), Ben Leinfelder (NCEAS), Matt Jones (NCEAS), Jing Tao (NCEAS)

5.2.2   Full report available at:
https://asm.lternet.edu/2009/workgroups/metrics-data-package-quality

5.2.3   <u>Executive Summary</u>: EML metadata contributed to the LTER NIS has not always been as usable it should be, and currently there is no mechanism for checking on the accuracy and structure of EML documents beyond simple schema compliance. This group assembled to consider adapting the "EML data manager library" for this purpose. LTER IMs would act as use cases and contribute their own experiences with automated reading of EML. Code would be written by LNO and NCEAS programmers. The goals for this group are to:
a) Establish a set of secondary metrics for LTER EML data package quality,
b) Recommend content for a report on data package quality (metadata and data) to be produced by the EML data manager library, and
c) Consider implementation strategies. These might include a quality report as another choice on the EML parser HTML page, or a shell script similar to that included with the EML parser.

Appendix:

- **IMC Governance Working Group (IMC GWG) Report**
- **GIS Working Group Report**
- **Controlled Vocabulary Working Group Report**
- **Unit Working Group Report**
- **EML Metrics Working Group report**
- **Network Information System Advisory Committee (NISAC) Report**

IMC Governance Working Group (IMC GWG) Report
Information Management Committee (IMC) Meeting September 2009

Full group discussion: The discussion of governance focused on decisions made by and communications between committees and boards across the LTER organization framework. A figure developed by the IMC GWG over the last year was shown an aid to facilitate discussion. Some current roles and practices were discussed:

- LNO recommends when working groups or sites have needs for resources and implementation that they contact NISAC as a way to approach the EB.
- Some NIS modules are developed by sites, while some are a whole network effort (PASTA); NISAC could play a larger role in reviewing and coordinating this.
- We need to be aware as we move from informal to semi-formal to formal in our practices, the strengths inherent to the informal/semiformal and that exploring similarities and differences make for stronger products
- A critical mass of sites may be useful for vetting
- Science working group projects go through an approval process with the EB, while LNO holds funds for the IMC, but how are projects funded? For now projects may be funded outside the IM. It would be useful for IMExec to have an annual budget so they can plan accordingly.
- Form of affiliation for projects may be useful to have, each with its' own criteria, a designation as to their relationship with LTER

    - LTER affiliated
    - LTER Endorsed

    And if part of LTER they have a designation as to their development when related to NIS, i.e.
    - NIS core, in production
    - NIS core, in development
    - NIS core, proposed
    - NIS module, in production
    - NIS module, in development
    - NIS module, proposed
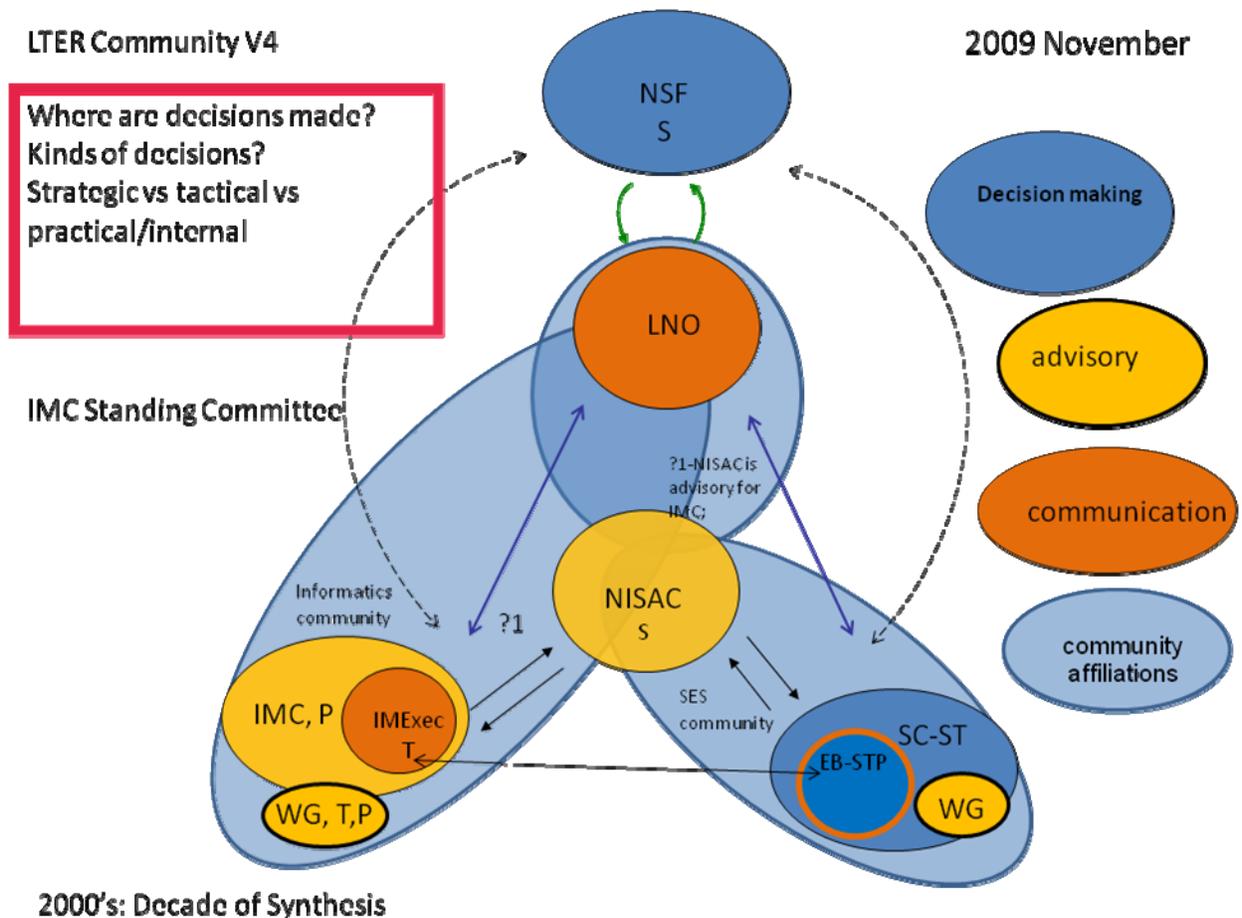
Please note that this diagram may be missing elements as the types of decisions and who makes them was not discussed in detail at the meeting or at any time by larger groups of information managers. However a need was identified to improve procedures for knowing who needs to be making decisions. Alternate models may need to be explored.

We focused on how IMC and IMExec interfaces with NISAC and the EB, as well as the LNO to facilitate and support progress on potential NIS modules, standards, and other projects related to IM. NISAC is well positioned to broker communication between IMExec and EB. The work of the controlled vocabulary working group was used as an example to explore not just who, but how decisions are made regarding the design, development, testing, support and adoption of IM or science driven CI "products". Working groups that have an idea or design for a product should speak with NISAC early on to get people thinking about how products can be developed to fit within a coordinated CI LTER NIS framework. This correspondence can happen in the form of a short report including the current status of the effort, useful ways it may be implemented, and some technical specifications. NISAC can use this information to propose how such tools may be incorporated into other projects. It was recognized that we have moved

from developing tools, standards, and best practices in informal ways, to using semi-formal consensus to advance efforts, and now may need more formal steps. During the design and development phase of any project NISAC may work closely with IMC, working groups, IMExec and LNO. At the point when resources are assigned from LNO a more formal process should be followed between NISAC and the EB.

We also discussed the partnership between IMC and LNO and how LNO resources are commonly available when LNO staff are serving as active participants and developers on working groups. In 2004, NISAC developed some requirements for potential NIS modules. It was recommended that this document should be revised. In addition, since 2004 network governance has changed so that EB now would now assign resources through LNO. This arrangement, however, means that technical proposals are not vetted through a group with the appropriate expertise. This procedure needs to be revised to handle proposals in a more informed way so the scope and ramifications of the project as well as its coordination with the larger network framework is taken into account. This may mean that in addition to the traditional flow of information from NISAC to EB that EB will ask for a review or advice from NISAC on issues relating to NIS and data. In addition there was wide agreement that potential NIS modules and other projects related to IM standards, protocols, best practices, and controls should address in written form a set of topics including but not limited to design, prototyping, testing, maintenance, update, site-level enactment, training, etc.

GIS Working Group Report from the 2009 All Scientist Meeting

Theresa Valentine, Andrews LTER   September 13, 2009

The following items were covered during the GIS Working Group Workshop Time at the 2009 IM Meeting:

1. GIS Standards for LTER Sites

    a. Proposed GIS Recommendations, question on how to move them forward, NISAC? IM Exec?

    b. The proposed recommendations were reviewed and revised. The working group will develop a formal proposal and present to IM Exec and NISAC. There was a lot of discussion during this session and under the governance session on how to proceed.

    c. Recommendations are as followed: September 2009 Edition
        1. All sites provide location data for the site headquarters into SiteDB
        2. All sites provide site boundary polygon to LNO and a boundary of Interest for Site (extending outside the site for DEM's, imagery, etc) (to be determined by the site)
        3. Basic Spatial Information Recommendations for each site:
            a. Require sites to place "basic spatial information" on-line
            b. Sites determine accuracy/resolution of data and document in metadata
            c. Document provenance of data
            d. Study site description information that is required for journal articles
            e. Include relevant metadata for all spatial data, including all required EML elements such as projection, coordinate system, and datum
            f. Digital Elevation Model (DEM)
            g. Imagery (Landsat TM scene)
            h. Research plot locations (based on GPS) of core research plots
            i. Roads
            j. Hydrography
            k. Structures
            l. Collect geo-referencing information for every research plot/site/observation location. Include as GIS file format or coordinate pairs (x, y locations).

        4. Spatial data metadata should be searchable with other site databases.
        5. Develop a centralized portal for users to search, display, and access all site spatial data.

2. GIS/Remote Sensing Training needs for LTER Sites

    a. GIS Data discovery training: how to find data and evaluate it for appropriate use.

    b. Higher end GIS training: (server and internet mapping applications, training that would work toward GIS Certification for site GIS personnel, modeling, and using the new Flex and Silverlight API's)

    c. Matching GPS data with remote sensing data (including LiDAR).

d. Google Maps training with IM's and using LTERmaps to help sites imbed the technology in their webpages (working on workshop proposal for this winter 2010).

e. Conduct a survey of web mapping tools and skills at LTER sites.

3. Other GIS issues/projects:

a. LTERmaps Phase 1 was demonstrated to the group. LTERmaps is a common interface to data about LTER sites through data in SiteDB along with site locations, site boundaries, gaging stations, and some EcoTrends data. The site uses a Google Map API. The Server subgroup has been working on implementing this technology since January, and is conducting a workshop during the ASM. The current demo site is http://gce-lter.marsci.uga.edu/public/gis/LTERmaps.html

b. Are their data transfer requirements/standards that we should document as best practices?

c. What happened to the spatial data workbench? What is the status of the project (http://www.lternet.edu/technology/sdw/)

d. The IM portal was discussed and the members of the GIS Working Group were asked to help populate and use the portal. There are several GIS related sections/projects.

e. Barbara Nolen agreed to track the Landuse/Landcover efforts that are evolving within the LTER Network.

4. ASM Workshops and GIS Lunch: The several GIS related workshops were discussed and participation encouraged. Several members of the GIS working group met for lunch on Wednesday.

Assignments:

1. Theresa Valentine to begin process of writing white paper proposal to formalize GIS Standards for LTER Sites.

2. Barbara Nolen to coordinate/track with LTER Landuse-Landcover efforts.

3. Need to follow up with IM Exec to forward GIS Training needs

4. LTERmaps team will apply for post-ASM funding,

5. Conduct a survey of LTER Internet mapping sites/skills (LTERmaps team)

6. Theresa Valentine to coordinate with emerging Spatial Analysis Committee (committee is being formed out of workshop http://asm.lternet.edu/2009/workgroups/lter-remote-sensing-data-information-and-coordination)

Controlled Vocabulary Working Group Report

IMC Meeting 2009: Notes by John Porter   13-Sep-2009

- See IM website for WG documents. DataBits article summarizes the problem.
- Steps taken: assembled list of LTER EML keywords, cross linked to other lists (NBII, GCMD, Metacat searches), edited to preferred form (kept track of synonyms), removed specific places and taxonomic names (better handled by other types of lists).
- Selected keywords shared with GCMD and NBII or used at more than one LTER site. Reviewed by IMC (removals and additions), voting by Survey Monkey. Edited list according to vote. If vote was close, went with current status.
- Resulting list of 640 words. 148 synonyms. 201 NBII keywords. 21 GCMD keywords. See website for list.
- Next steps. Is additional editing required? Who decides if it is an official LTER list? What procedures for subsequent editing? Who should manage list database (term, scope, definition, synonyms)? Analogous to unit dictionary.
- Tools. Auto-complete search tool (done). Auto-complete keywording tool (web page demo). Backend database for auto-completion resides at LNO. Dynamic pick list. But not enforced.
- Update document keywords tool. Advanced search tool.
- Hierarchies. Keywords are most useful when they can be tied to other keywords. Taxonomies, poly-taxonomies, thesauri.
- Term, broader term, narrower term, synonyms.
- Existing KNB browse hierarchy is rather limited. See LNO Metacat.
- Mapping of KNB categories to NCEAS sort list (higher categories).
- Assist in searches by offering an option to go up from keyword to broader concept and increase number of hits.
- Scope of problem is limited by relatively small number of keywords (640).
- Library community recommends providing context for words with different meanings (e.g. "head").
- What to do with other lists? Gazetteer for place names? Taxonomic list and tools?
- Original list was about 21,000 words.
- Tracking keyword searches in Metacat. Problem with misspelled words. Good mix of words.
- Mark Schildhauer has worked extensively with other keyword lists. Most are more narrowly focused. LTER list is quite broad. Important for different efforts to keep in touch. Corinna could report on SONET activities. NASA is developing SUITE (upper level). Margaret could report on SemTools project, which will make Morpho a more competent data management system. Hierarchy = clear subsumption hierarchy. Other relationships better described with OWL.

- Tools to assist with legacy EML. Auto-completion tool will handle new EML documents. Tool for existing EML could provide a useful first cut. More sophisticated tools could provide suggestions based on document content. Semantic annotation. Hook to parents, children, and related terms might be useful.
- Educating PIs on using the list will be important.
- Are there datasets not discoverable with current keyword list? Identify "naked" datasets with no LTER keywords or synonyms. Some keywords will be site specific.
- This list is not exclusive. An EML file may contains others as well. E.g. use <keywordList> to identify LTER keywords. Current name-spacing could be relaxed beyond "theme" and "place" to provide more flexibility. Tools could target specific keyword lists.
- EcoTrends preceded this effort but would like to take advantage of it. Sometimes title does not match dataset, keywords could help. Synonym ring could help.
- Multi-word keywords (e.g. primary production) are more difficult to reconcile with titles.
- Subtle problems: Primary production is a rate, primary productivity is a thing. Does "bug" mean "true bug" or "insect"?
- How are tools and technologies tied to organizational routines? E.g. how to get PIs to use the list? How will users be trained to think in categories? Search engine will assign higher level categories (not scientist). Auto-completion tools will help enforce standard spelling.
- Pursue poly-taxonomies based on Barbara's list. Probably no more than 4 levels. AND has 3-level hierarchy. Should hierarchical structure be displayed? Use expandable list (not a single long list).
- Develop synonym list.

Unit Working Group Report

2009 LTER Information Management Meeting

The LTER IMC Unit Working Group is an endeavor that began in 2004. It has developed into a Unit Initiative with three major thrusts: a Unit Dictionary, a Unit Registry, and a Unit Best Practices guide. Work with units across the LTER sites began with a cross-site project that acknowledged and discussed the heterogeneity of units used with biological datasets. This effort culminated in 2005 with a demonstration of a web interface that provided access to and made visible a community designed unit dictionary. Subsequent work by a Unit Task Force focused on developing a deeper understanding of and broader experience with the heterogeneity of custom units as defined by EML. These efforts demonstrated the need to review and document community interpretations of how to initiate sense-making activities relating to site-level differences in the use of units.

The Unit working group session at the 2009 IM meeting was led by Lynn Yarmey (CCE/PAL) and Mason Kortz (CCE/PAL). In attendance were:

      Hap Garritt (PIE)

      Dan Bahauddin (CDR)

      Suzanne Remillard (AND)

      Florence Millerand (CCE/PAL)

      Karen Baker (CCE/PAL)

      Dave Balsiger (NTL)

      John Campbell (HBR)

      Helena Karasti (FinLTSER)

      James Conners (CCE/PAL)

      Ben Leinfelder (NCEAS)

The session began with a report on the progress of the IMC Unit Working Group over the past year. Two major efforts were defined as part of a LTER Unit Initiative: the Unit Dictionary project, comprising the set of units in use by the network and the best practices that cover them; and the Unit Registry, the application used to access and manage these units.

Lynn Yarmey presented the current version of an LTER Unit Dictionary Best Practices developed by the group over the last year. Prior to IMC members, she sent out and posted online the current version of the Best Practices. Discussion took place for each practice and alterations suggested by the group were recorded. This was followed by a more general discussion on the authority of the best practices. It was agreed that units should not have to meet all of the best practice standards to be entered into the Unit Dictionary and Registry, but

ultimately a vetting process will be developed to further encourage and inform network standardization of unit practices.

Mason Kortz gave a demo of the Unit Registry prototype developed at the PAL/CCE site. The Unit Registry will provide services and interfaces for querying and managing the Unit Dictionary. The demo showed the ability to retrieve units, quantities, and scopes from the database via a web service interface. These elements were combined to create a demonstration EML unit validation service. Discussion was held after the demo to answer questions about the current service and gather suggestions for new features and possible site- and network-level use cases for the software.

The working group closed with a request for volunteers for both Unit Working Group efforts. Volunteers interested in focusing on content will help review current units in the Unit Dictionary for compliance with the best practices. They will also revise the best practices document, adding any necessary clarification. Volunteers interested in development aspects of the Unit Registry will work on implementing use of the dictionary and registry in site applications, providing feedback and bug fixes. Additional volunteers who did not attend the working group signed up during the IMC meeting. The volunteer lists are:

Content Volunteers:

 Hap Garritt (PIE)

 Dan Bahauddin (CDR)

 Suzanne Remillard (AND)

 Dave Balsiger (NTL)

 Margaret O'Brien (SBC)

 Nicole Kaplan (SGS)

Development Volunteers:

 Ben Leinfelder (NCEAS)

 Sven Bohm (KBS)

 Kyle Kwaiser (UMBS)

 Ken Ramsey (JRN)

 Adam Skibbe (KNZ)

EML Metrics Working Group report

Proposed Work Group: Metrics and reports for EML data package quality
September 17, 2009 (ASM, Estes Park)

Attendees: Margaret O'Brien (SBC), Emery Boose (HFR), Dan Bahauddin (CDR), James Brunt, (LNO), Mark Servilla (LNO), Duane Costa (LNO), Mark Shildhauer (NCEAS), Ben Leinfelder (NCEAS), Matt Jones (NCEAS), Jing Tao (NCEAS)

The primary quality standard for EML documents is XML schema compliance and the EML parser. Schema compliance is usually enforced by the editor used to create the document or checked by the EML parser, which also checks that rules for EML ids and references have been met. Experience using EML metadata contributed to the LTER NIS to automatically read and make use of data entities indicates that a significant fraction do not have metadata of sufficient quality for this use. The primary contribution from LTER sites to the NIS is data sets, which are intended to be used in cross-site synthesis projects. Clearly, any automated use of EML in the NIS will require a higher level of metadata and data quality.

The EML data manager library was created to read and parse EML metadata documents, then to download the data entities and store them as tables a relational database. It can also query those tables using SQL-like constructs. For a table to be ingested, its metadata must be accurate (not simply valid EML) and its format must be clean, consistent and match the metadata precisely. So the data manager library can be used to create the next level of quality control checks for EML datasets and their tables.

The goals for this group are to:

1. Establish a set of secondary metrics for LTER EML data package quality,

2. Recommend content for a report on data package quality (metadata and data) to be produced by the EML data manager library, and

3. Consider implementation strategies. These might include a quality report as another choice on the EML parser HTML page, or a shell script similar to that included with the EML parser.

Initially, the quality reports can be used to
1. Inform the dataset contributor about the content of the data package, and indicate whether data are of sufficient quality to be machine-readable. XML repositories have no quality standards beyond basic XML schema and (in the case of Metacat) EML compliance, so a data package that fails these secondary quality metrics can still be uploaded or harvested. However, a dataset contributor should be aware that the usefulness of the dataset will be limited.

2. In the LTER context, collating reports can produce a list of failure modes for LTER metadata and data entities. Such a list could provide input for the design of specific tools for data providers, or help identify gaps in an LTER Site's IM system. A Site requesting supplemental funding for its IMS could use the reports as part of the proposal justification.
Additional details about the dataset may be reported at some later date, e.g., basic stats, ranges, frequency distributions, and also may be compared to metadata values. A preliminary figure showing some possible failure modes is attached.

Figure 1. Possible failure modes and reports on data quality or content for an EML data package being read by the data manager library. The boxes labeled "why not?" might indicate places where a site's IM system could be insufficient to produce NIS-ready EML.
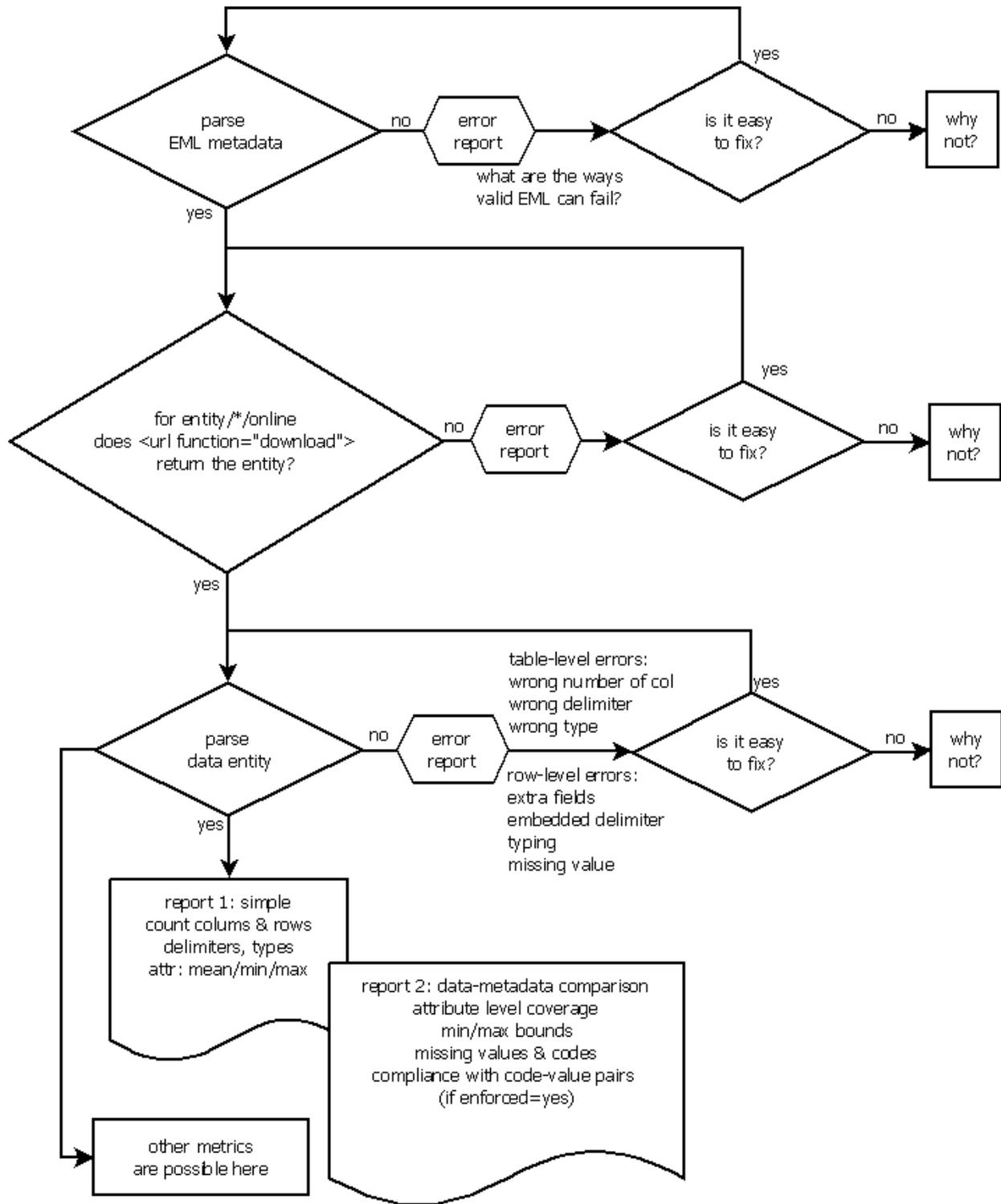
CI Implementation Working Group – ASM 2009

LTER

Network
Information
System
Advisory
Committee

Wade Sheldon
NISAC Co-Chair
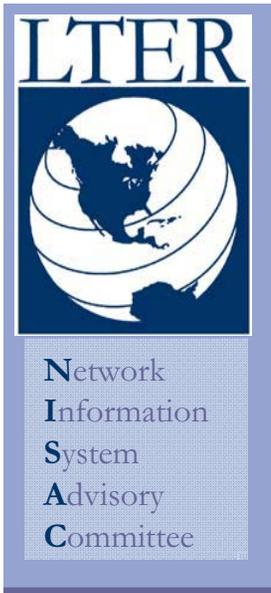
CIIP Process Overview

LTER

- NISAC tasked by LTER EB to draft a specific plan to begin implementing CI components of Decadal Plan (aka CI Strategic Plan)

- Plan needs to include:
  - Tasks (what and who)
  - Priorities (when)
  - Funding Source (how)

- IMC briefing presented via VTC (Jul 2008) – on IM web site

- NISAC finalizing draft based on Spring 2009 meeting

- Availability of ARRA funding stepping up the time scale

- Good time to pursue low-hanging fruit, make some quick gains

CIIP Tasks (*DRAFT*)

LTER

Network
Information
System
Advisory
Committee
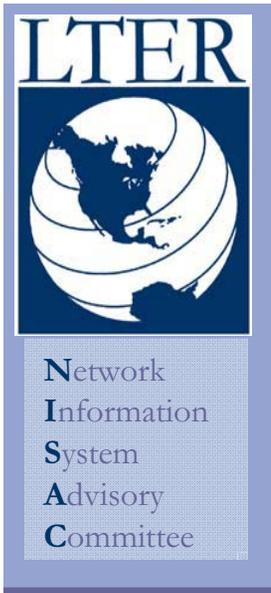
- Initiative 1: Build community-based services and a service-oriented architecture (SOA)

  - Near-term tasks
    - Develop web service interfaces for network databases
    - Migrate and ClimDB/HydroDB to LNO, update web services, ensure CUAHSI interoperability
    - Identify and prototype an auditing/notification service for network resources

  - Mid-term tasks
    - Identify and deploy federated authentication/ Single Sign On (SSO) and security technology
    - Evaluate and select schemas for web service data exchange (beginning with EML)
    - Prototype web service wrappers for site systems
    - Identify and prototype middleware for connecting applications with distributed data
    - Identify and prototype a network resource discovery/ management service

  - Long-term tasks
    - Identify and prototype integrated applications based on web services
    - Develop and deploy Point-of-Presence nodes at the sites (standard computer configurations and software stacks)

CIIP Tasks (*DRAFT*)

Network
Information
System
Advisory
Committee
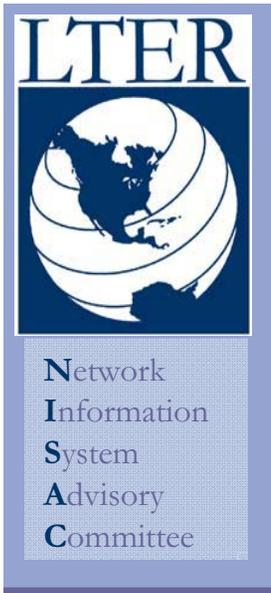
- Initiative 2: Build CI capacity to increase data acquisition, management, and curation at the site level

  - Near-term tasks
    - Improve standardization and quality of LTER EML documents
    - Complete and adopt controlled vocabulary for keywords and EML unit dictionary
    - Standardize automated direct access to site data
    - Identify and evaluate sensor network management approaches being developed by sites and EONs (put together document of recommendations)

  - Mid-term tasks
    - Develop standardized attributes (names, scale, units) for common dataset parameters (as in climate standard)
    - Evaluate, develop automated QC procedures for high volume data
    - Define standards for QA/QC and missing value annotation in site data
    - Identify or develop common data models, data warehousing approaches and best practices for site data

  - Long-term tasks
    - Evaluate technology for automated metadata and data capture in the field (e.g. technology for replacing paper forms with PDA/GPS)
    - Identify common high-impact data sets that all sites should provide to support network research agenda (land use, PDI, PET, NPP, chemistry, LIDAR, demographic and socioeconomic data, GIS and remote sensing)

CIIP Tasks (*DRAFT*)

LTER

Network
Information
System
Advisory
Committee
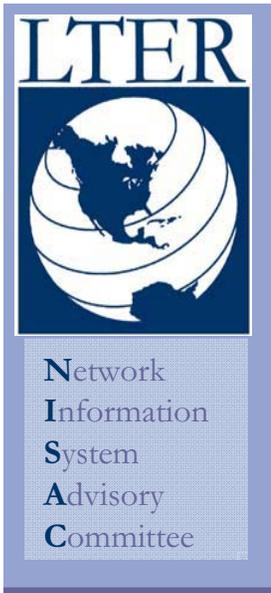
- Initiative 3: Build CI capacity to increase data discovery, access, and integration

  - Near-term tasks
    - Implement and deploy "full" PASTA architecture for "key" LTER data
    - Leverage other data networks (such as CUAHSI, GEON) to support integration with LTER data

  - Mid-term tasks
    - Identify and prototype observational data model for standardizing LTER data
    - Identify and prototype persistent identification system for accessing LTER data and metadata (e.g., DOI or LSID)
    - Evaluate use of attribute-based and other data descriptive specific ontologies

  - Long-term tasks
    - Prototype EML-based framework for exploring "Dataspace" type data discovery and integration
    - Design and prototype automated systems for QA annotation and classification of LTER data
    - Evaluate warehousing approaches vs. distributed queries (different approaches may be needed for different classes of data)

CIIP Tasks (*DRAFT*)

LTER

Network
Information
System
Advisory
Committee

- Initiative 4: Build CI capacity to increase modeling and analysis activities

  - Near-term tasks
    - (none identified)

  - Mid-term tasks
    - (none identified)

  - Long-term tasks
    - Develop standard for documentation of models and model inputs and outputs
    - Explore existing systems for documenting and storing models to support re-use (model description, analyses, suitability)
    - Develop (or leverage) shared repository for model code and test-bed datasets
    - Evaluate Grid services to support distributed models and collaborative model development
    - Establish CI requirements to support a potential modeling and analysis center (e.g. decide among distributed versus centralized approaches)

CIIP Tasks (*DRAFT*)

LTER

Network
Information
System
Advisory
Committee

- Initiative 5: Build capacity to increase collaboration

  - Near-term tasks
    - Survey technologies used by recent working groups (Gragson, Collins) to evaluate functionality and effectiveness
    - Deploy collaborative environment for IM, science and education working groups
    - Deploy collaborative software development infrastructure (forums, code versioning system)

  - Mid-term tasks
    - Explore strategies for increasing bandwidth available at field sites (Last Mile Connectivity)

  - Long-term tasks
    - (none identified)

CIIP Tasks (*DRAFT*)

LTER

Network
Information
System
Advisory
Committee

- Initiative 6: Integrate cyber-infrastructure into social-ecological research, education, and training

  - Near-term tasks
    - Develop and conduct workforce education and training for scientists and IMs

  - Mid-term tasks
    - (none identified)

  - Long-term tasks
    - Support technologies for providing remote education and training (web-casting, field-based experience)
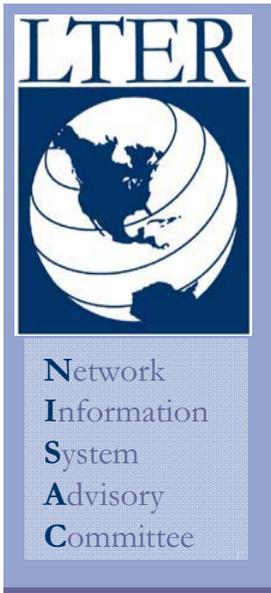
CIIP Tasks (*DRAFT*)

- Initiative 7: Collaboration and Integration with other Observatory Networks

  - Near-term tasks
    - Support LTER IM and scientist participation in CI collaborations (workshops, standards-setting organizations, training) - establish responsibilities for reporting, staying in contact
    - Conduct broad EIM (Environmental Information Management) meetings that engage CI partner organizations
    - Develop proposals based on LTER CI partnerships
    - Collaborate with GSC to develop standard for linking "omic" data with environmental observations

  - Mid-term tasks
    - (none identified)

  - Long-term tasks
    - (none identified)
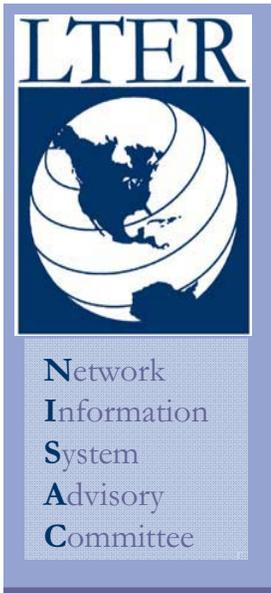
IMC2008 CI Working Group

- LTER needs a CI "framework" to guide activities
    - Decadal Plan is a starting point (broad goals)
    - NISAC-lead CI implementation plan can provide specifics
    - Needs to be a living framework, revised based on experience, vision

- IMs need to stay informed about network CI initiatives - need more discipline on keeping up with activities (LTER, IMC web)

- CI initiatives need to include well defined interfaces and exchange standards to ensure broad use
    - IMs need to participate in defining use cases, design requirements
    - LNO NIS developers need specifics!

- Some sites limited in personnel and IT hardware - need some targeted investments to bring in staff and hardware

IMC2008 CI Working Group

LTER

Network
Information
System
Advisory
Committee
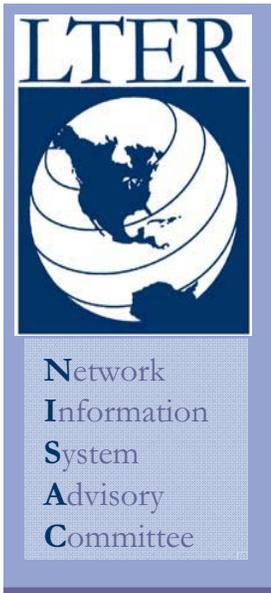
- IMC CI Priorities over the next year
  - Web services interfaces to LNO databases
    - Enable leveraging on site websites
    - Synchronizing automatically between site and LNO (B2B)
    - Links between databases (personnel, sites)
    - Prototypes, early milestones, broad input critical
  - Generalized quality control tools for streaming data
    - Need ways to organize and house high volume data
    - Need to emphasize shared solutions, models
    - Need tech transfer, training on IM-developed and commercial tools
  - GIS infrastructure (Geoserver, storage, shared archive)
- Money was requested in LNO budget to support visits to and from LNO
  - Initially omitted (flat funding)
  - Now back in budget, plus ARRA funding

IMC2009 Working Group

LTER

Network
Information
System
Advisory
Committee
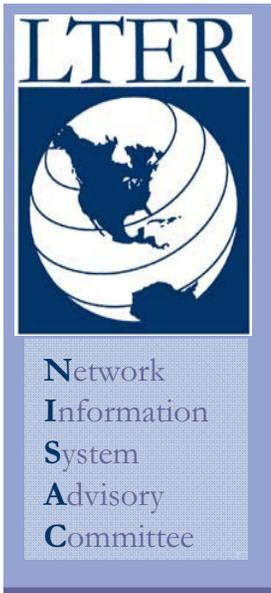
- Web Service Interfaces
  - Steps forward
    - Identify specific functional requirements
    - Adopt/develop exchange formats/schemas, methods, end-point targets
    - Identify resource requirements, time-lines
    - Develop prototypes
    - Vet and promote implementation
  - How to plan and coordinate
    - Collaborative working group model (ala ProjectDB)?
    - IM designates to work with LNO NIS, develop RFCs and prototypes, seek IMC input/review?
    - IMC provides LNO NIS with needed specs, serve as adopters, testers?
    - ???
  - Brainstorming candidate use cases, needed interfaces/APIs
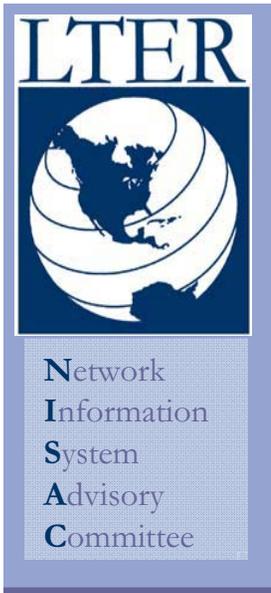
IMC2009 Goals

- **High Volume Data Management, Q/C**

  - Does this fit within ARRA funding, supplements?

  - How to define specific needs?

  - Candidate training topics?

  - What coordination resources needed?

    - IMC web site forum, projects

    - Working group

    - Collaborations outside LTER
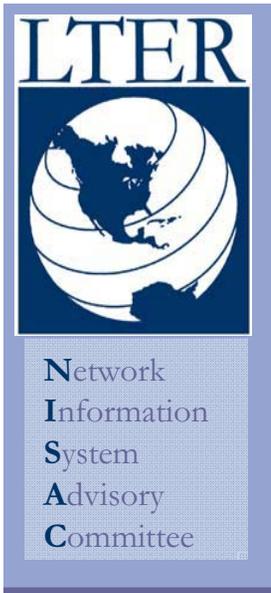
IMC2009 Goals

- **GIS Infrastructure Development**
  - Does this fit under current GIS working group scope?
  - What additional planning/coordinating resources most needed
  - How do we identify priorities?