NISAC Annual Report

May 1, 2012

Contents

Summary

This report servers several purposes. It summarizes the issues addressed by the NISAC from May 2011 through April 2012; documents the committee's evaluation of the LTER network information systems; and provides recommendations to the Executive Board (EB) for further action. For the convenience of the EB, we provide a succinct list of recommendations in this section. However, we encourage the EB to review the entire document to understand the rationale behind these recommendations and to gain an appreciation of the efforts and progress made by those responsible for the NIS. Finally, we remind the EB that a second document, ***LTER Data Accessibility: Barriers and Solutions***, is provided per the request of the EB. This "white paper" has an additional set of recommendations relevant to network data access policy, metadata, and data discoverability.

NISAC recommends that:

1. PASTA be used in an ongoing multi-site science effort (p. 4);
2. a comparison be made between the list of data sets that sites have agreed to publish, as indicated in renewal proposals, and data sets that can be found in MetaCat (p. 6);
3. the LTER Executive Board contact Dr. Blood (NSF) to clarify the status of funding needed to publish legacy datasets (p. 7);
4. the Executive Board either create the Data Council or assign its responsibilities to another entity (p. 7);
5. the Executive Board review and revise objectives for community-wide efforts to integrate data management systems in light of the recommendations of the 30 year review and provide guidance to NISAC on how to evaluate this objective in the future.
6. the IMExec proposal on Promoting Data Sustainability by supported (p. 8);
7. the LTER Network host (i.e., place on a server, maintain backups) but not curate (e.g., transform the data from spreadsheets to data tables, prepare metadata, and user support) LINX and LINXII data (p. 9).

NISAC members

Co-Chairs

Paul Hanson (NTL)

John Porter (VCR)

Members

James Brunt (LNO)

M “Gastil” Gastil-Buhl (MCR)

Chuck Hopkinson (PIE)

Bill McDowell (LUQ)

Will Pockman (SEV)

Linda Powell (FCE)

Suzanne Remillard (AND)

Mark Servilla (LNO)

Bob Waide (LNO)

Libe Washburn (SBC)

PASTA evaluation

Summary statement
Overall, NISAC is very pleased with the progress on PASTA. An important process in the development of PASTA was to incorporate feedback from the user community via Tiger Teams, including several NISAC members. Due to the design and development goals guided by this input, NISAC is confident that PASTA is well poised to support the data sharing requirements of network science.

Background
After receiving ARRA funds from NSF to implement the Network Information System's (NIS) PASTA Framework, the first task was to write a draft Operational Plan and then have the draft reviewed. Following completion, the draft Operational Plan was reviewed in Fall, 2009. Based on the review, the Operational Plan was modified and released in February 2010. It was this plan that the NISAC reviewed during its March 2011 NISAC meeting and its April 2012 NISAC meeting. Here we present comments made during the April 2012 meeting.

NISAC response to progress report presented by Mark Servilla

1. NISAC science members supported the idea of developing an easy-to-use data access portal that does not require knowledge of the underlying data handling system.

2. The quality report produced by PASTA upon submission of data packages will be of value to by Information Managers (IMs). Processes leading to the quality report eliminate poorly described or otherwise deficient data sets from being stored. This, for example, will allow acquisition of data from all sites and guarantee that there are data behind each link.

3. Site data curation will remain essential. PASTA will leverage capacity of existing site information management resources. Site-based information management will remain a crucial aspect of site-based and network-level science activities. A possible misconception is that IMs will have nothing to do following PASTA completion.This is not the case.

4. PASTA is on track to assign digital object identifiers (DOIs) to data sets in the second phase of development (Summer 2012), which will be beneficial in many ways. DOIs provide end users with an unambiguous and persistent identifier to digital objects available on the Internet - once assigned to a digital object, the DOI will not change. In addition, DOIs are recognized by publishing industry for

referencing digital objects.

5. NISAC discussed a preliminary “roll-out” plan with two components: (1) NISAC recommends using PASTA in an ongoing multi-site science effort; (2) LNO plans a “market research” effort to evaluate PASTA’s prototype Data Portal user interfaces, which NISAC endorses.

6. Some non-LTER users of PASTA have emerged. For example, the Paleoecology Observatory Network indicated that PASTA is the most appropriate information system for their data management needs. This demonstrates the potential for broader use of PASTA.

7. So far, PASTA handles tabular data. It is planned that the next phase will handle other types of data, which will be an important advancement.

Review of IMExec proposal on *Promoting Data Availability*

IMExec has developed a document containing a table called ‘Promoting data availability’ that identifies and prioritizes activities that will lead to better data access. The activities are linked to general funding activities, SIP milestones and entities responsible for accomplishing these goals. NISAC endorses this document as a framework for applying available funding towards tasks that will lead to increases in data availability.

SIP review and response

The following evaluation addresses the five Strategic and Implementation Plan (SIP) items the committee thought most important. The numbering corresponds to the numbers in the Information Management section of the SIP.

1A. Fully document site data in accordance with Network standards

A. Document and create rich EML for data collected and available at each site

The SIP calls for sites to fully document existing site data, including the development of rich EML documents for each data set, by the end of 2012. Most sites have made significant progress towards or achieved this objective. A few sites still face significant challenges to reach this objective, either because they consider this task a low priority or because site resources are insufficient to meet the challenge. Sites should elevate the priority of the

actions required to meet this objective to assure that data are ready for ingestion into the Network Information System once the PASTA framework becomes operational in 2014. Sites should conduct an inventory of existing and planned data sets and develop a strategy for documenting data sets that still lack rich EML. These inventories will provide a baseline to measure and report progress.

For those sites facing barriers attributable to lack of resources or expertise, the LNO and Information Management Committee should develop means to transfer expertise from knowledgeable information managers to sites requiring assistance. Sites whose Information Managers provide expertise should be compensated for the loss of time, either by the site requiring assistance or the LNO. Emphasis should be on training information managers to become self-sufficient in metadata creation. Clear goals such as increased numbers of well-documented data sets should be established for each site receiving assistance.

1B. Develop quality control standards for LTER data that meet needs identified by projects requiring data synthesis across the LTER Network, and implement these standards at each site to create a high level of confidence in LTER data

A. Define general QA/QC approaches

During 2011 LTER a SensorNIS workshop was held at Hubbard Brook focusing on quality control and assurance of streaming sensor data. Additional workshops are planned for May 2012 addressing aspects of sensor data and QA/QC. Progress has been slowed in achieving this goal because few funded projects require synthesis across the LTER Network and because no entity has yet to assume responsibility for developing priorities among the many possible synthetic data sets. The StreamChem DB project is one exception, and could provide an opportunity to establish QA/QC approaches for certain classes of data. NISAC should recruit a domain scientist familiar with the plans for development of StreamChem DB and provide oversight and encouragement for the implementation of QA/QC standards developed by this project across the Network.

1C. Increase the amount of data shared by LTER sites.

A. Obtain a commitment from each site to abide by LTER data sharing policies

The LTER Network has a long-established policy of open sharing of data whose goal is to make data available to the broad scientific community and the public in a timely manner. The policy and certain exceptions are spelled out clearly in documents approved by Network governance. However, information on the degree to which sites conform to this policy is not

available, and hence its success cannot be evaluated. The LTER Network should engage an outside entity to evaluate accessibility of LTER data through 1) site web sites, and 2) the LTER Metacat Data Portal. To begin this process, sites submitting renewal proposals in 2012 should make available tables of site data sets and URLs that were submitted with the renewal proposal. Existence of these data sets on site web sites and in the MetaCat will be confirmed through an outside entity, with reports going to each site's Lead PI and to NISAC. NISAC will establish a baseline of data available through the MetaCat, and LNO will gauge progress in increasing the number of data sets available on an annual basis.

By May 2012, NISAC will produce a white paper describing barriers to data sharing and potential solutions to these barriers.

1E. Digitize or parse, and clean legacy data into an accessible electronic format

A. Identify and prioritize datasets

In the absence of the proposed Data Council, the Data Synthesis Project has proceeded through the efforts of the ad hoc Synthesis Data Committee, the LNO, the IMC, and three sites (BNZ, CWT, SGS) that have agreed to participate in the project. A Request for Proposals (RFP) was prepared by LNO and reviewed by participating sites and NSF. This RFP is now ready for distribution through the University of New Mexico Purchasing Department, but is awaiting final approval from Dr. Elizabeth Blood of NSF. The LNO has agreed to provide some support to cover the costs of site participation in the project. NISAC recommends that the LTER Executive Board contact Dr. Blood to clarify the status of the project.

The Executive Board has yet to create and populate the Data Council. Hence, the task of identifying priority data sets for the Data Synthesis Project and other derived data sets has yet to occur. The Executive Board should either create the Data Council or assign its responsibilities to another entity.

3B. Develop and deploy network-wide data collection, storage, and delivery operations that promote Network synthesis and the creation of data legacies and open access to LTER data products

A. Continue to implement the Operational Plan for the LTER Network Information System

The LNO software development team, Mark Servilla, Duane Costa, and James Moss aided by Tiger Teams made up of LTER scientists and information managers, has met the milestones described in the Operational Plan and will be ready to roll out the initial prototype of the PASTA framework for the LNO

mid-term review in May 2012. A separate report will provide more details of progress.

4C. Mentor and facilitate development of integrated data management systems by environmental observatories that are collecting similar types of information, including but not limited to ULTRAs, LTREBs, OBFS sites, the National Phenology Network, the Genomics Standards Consortium, the Critical Zones Observatory program, the Ocean Observing Initiative, the Arctic Observing Network, and Earthscope.

A. Communicate our metadata standards to other networks

Responsibility for this task is not clearly assigned in the SIP. However, the LNO has continued to communicate metadata standards to various other networks and entities, including NEON, projects funded by the Macrosystems Biology program at NSF, the Organization of Biological Field Stations, the Critical Zones Observatory program, the Genomics Standards Consortium, and the former National Biological Information Infrastructure. The Executive Board should clarify where responsibility for this task resides.

B. Crosswalk between our metadata specifications and other specifications

This task was completed by the LNO before NBII ceased to exist in 2011. Currently in progress is the fgdc to eml crosswalk work being done by the GIS working group.

C. Provide leadership to develop, initiate and continue community-wide efforts to integrate data management systems.

The Executive Board needs to review and revise objectives in light of the recommendations of the 30 year review and provide guidance to NISAC on how to evaluate this objective in the future.

LINX data and LTER

CHARGE: Evaluate the request to have the LTER Network provide curation for the data from the LINX and LINX II experiments on the fate of nitrogen in streams. NISAC should provide recommendations regarding 1) whether the LTER Network should undertake curation of the LINX data, 2) if so, what entity within the network should be tasked with doing so?, and 3) what resources are likely to be required to complete the task of adding the data to the LTER Data Catalog in a PASTA-ready form?

The data from the LINX and LINX II cross site experiments constitute an excellent example of why planning and resources need to be dedicated to information management at the beginning of a project, rather than trying to “rescue” the data at the end of the project. The data currently reside is a series of non-uniform, cross-linked spreadsheets that would require expensive, time-consuming and potentially error-prone transformations to fit into tabular data formats.

NISAC examined sample copies of the spreadsheets and conferred with some LINX leaders regarding their willingness to undertake the transformations needed to go from esoteric spreadsheets into data structures, such as data tables. They indicated that they were interested in seeing the data preserved, but did not have the resources or interest in undertaking the transformation of the data themselves. As a stopgap measure, the spreadsheets, along with existing documents associated with the analysis, can be combined into a single .zip file to prevent further deterioration of the data as files are lost. However this is a short-term solution because the software required to read the spreadsheets may be difficult or impossible to obtain in the future (e.g., by 2025).

NISAC therefore recommends that the LTER Network host (i.e., place on a server, maintain backups) but not curate (e.g., transform the data from spreadsheets to data tables, prepare metadata, and user support) LINX data. We recommended to LINX researchers that they contact some knowledgeable LTER Information Managers to discuss options for providing the data, which might range from simply providing a .zip file containing the spreadsheets, along with whatever documentation is available, to processing the spreadsheets to create conventional datasets. They should explore either contributing the “Data Appendix” associated with the LINX II Nature paper to the LTER system or perhaps even transforming a copy of that material into formal datasets within the LTER system.