1-22-2010

LTER Cyber-Infrastructure Plan External Review

Long Term Ecological Research

LTER CI Plan External Review

January 21-22, 2010

Bruce E. Wilson (wb5)

Review Panel:
Terry Benzel: University of Southern California, Information Sciences Institute
Randy Butler: University of Illinois, National Center for Supercomputing Applications
Chris Jones: University of California, Santa Barbara
Michael Piasecki: Drexel University
Robert J. Sandusky: University of Illinois at Chicago
Bruce Wilson: Oak Ridge National Laboratory and the University of Tennessee

Summary

The Cyberinfrastructure / Networked Information System external advisory panel, as identified above, conducted a review of the Long Term Ecological Research (LTER) Network Office (LNO) Network Information System (NIS) Operational Plan (Draft 2, dated 12-18-2009). The operational plan supports the larger goals of the LTER and the LNO across four thematic areas: synthesis, cyberinfrastructure, core services, and development and outreach. This panel's charge was to focus on the cyberinfrastructure (CI) tasks, particularly those that are funded by ARRA sources, which comprise the majority of the tasks and the majority of the expenditures specified in the operational plan. The CI tasks are intended to provide innovative facilities to enable synthesis and network-level research -- enabling the creation of new knowledge by using data collected at multiple LTER sites. As such, the CI tasks are intimately related to the science, communication, and outreach tasks identified in the operational plan, and the advisory panel evaluated the CI plans in the context of that larger whole. In particular, development of a robust and effective NIS depends upon the ability of the LNO to motivate participation of personnel at the LTER sites; as a result, many portions of the operational plan describe the nature and extent of communication and collaboration needed to design, develop, test, and deploy an NIS that meets the needs of the LTER sites and scientists.

Our summary conclusion is that the operational plan is sufficiently comprehensive, organizes the sixteen major project tasks coherently, and scopes the effort at a generally appropriate level of detail. The tasks detailed in the operational plan are necessary for achieving the overall objectives, and the plan is realistic, albeit somewhat ambitious, in terms of the scope of work to be accomplished. The panel also has developed a set of recommendations, noted below and expanded upon in both the response to LTER questions section (page 3) and in the Additional Discussion section (page 7).

The advisory panel has grouped its recommendations into two groups: critical priority and high priority. The critical priority items have the potential, if not addressed, to delay the project schedule and / or impact the ultimate effectiveness and impact of the NIS. High priority items have the potential to reduce project / LNO / LTER effort in the long run or increase the impact and effectiveness of the NIS.

Critical priority:

- Ensure the engagement of the LTER scientists and information managers in the NIS development efforts. We recommend using LTER resources and potentially seeking additional funding to ensure the active engagement of the Information Managers and site scientists in the further design, development, and testing of the NIS. This recommendation includes developing and using an active communication plan for the NIS development within LTER.
- Document the software security approach to be utilized in the NIS. While LTER has significant experience in distributed software development and implementation, the PASTA architecture creates new security issues, particularly in the context of changing Internet security issues, and dramatically increases the importance of a consistent security approach across the LTER sites. We recommend that effort be devoted to documenting those approaches to ensure consistency within the NIS development, particularly in light of new staff likely to be utilized in NIS development, the increased interdependency of LTER sites, and the potential for community contributions and integration with the efforts of other projects.

High priority:

- Ensure transparency and wider community involvement. This expands the communication plan task to communicate the NIS developments to the wider community. We specifically recommend that LNO be explicit in communicating recognizing that this software will be developed as open source, and we recognize that the LNO this LTER has followed this approach in much of its past work.
- Ensure adherence to, acknowledgment of, and openness to standards. This includes both standards for web content accessibility, data storage formats, and data exchange protocols. This project should be intentional about identifying and using those key standards that will enable the necessary interoperability, which in turn will enable the target integrative and synthetic science.
- Implement a more active risk assessment process. LNO asked the panel if there were risks not identified in the operational plan. We note that very few risks are explicitly identified in the operational plan, though a number of risks were clearly identified in the discussions and presentations around the plan. Given LNO's experience, an extensive analysis and monitoring of risks will not be value-added, but documentation and monitoring of the identified risks is an appropriate investment.
- Identify key metrics to ensure that the project is on track during the entire cycle and to identify materialization of risks as early as possible. Many of the metrics identified were trailing metrics of success, important for measuring ultimate project success, but of less value in project management and risk mitigation. We recommend that metrics be identified and monitored which measure a) internal and external community engagement and b) progress against baseline.
- Ensure that the Application Programming Interface (API) is as rich as the Graphical User Interface (GUI). This is a generally accepted best practice, inherent in concepts of a Service Oriented Architecture (SOA), and implied by a number of aspects of the NIS design. We recommend that LNO be more explicit in this aspect of the NIS design, to ensure that the NIS is usable both by people and by other software tools.

Response to LNO Questions

LNO posed seven questions for the review panel to specifically consider in responding. Our answers to these questions are provided below. The discussion lead noted for each question is the primary author of that section, drafting a response that represents both that individual's particular expertise and the consensus of the review panel.

1) Are the operational steps presented adequate and in keeping with projects of similar scope and size to address the goals of the project?

(Discussion Lead: Bruce Wilson) Overall, the operational steps are adequate and in keeping with projects of a similar scope. The operational plan represents considerable thought and the experience of the LNO staff in operational ecological informatics. Given the uncertainties inherent in cyberinfrastructure development for a research environment, the general level of detail is appropriate, provided that some additional attention is devoted to metrics for project execution, community engagement, and management of scope and schedule risks. From an internal project management perspective, some additional detail will need to be developed for the iterations during project execution, as a natural part of the proposed development methodology (Rational Unified Process). This is not intended to require additional task detail in the operational plan but rather as a recognition of steps LNO intends to take and as a risk management strategy related to project execution risks and risks associated with the transition to operations.

2) Are the allocated resources adequate to accomplish the project as outlined?

(Discussion Lead: Bruce Wilson) The allocated level of resources is reasonable for the project scope, as outlined, although the scope is aggressive. Strategic and effective collaboration with other projects, management of scope, and appropriate use of standards will be important factors to ensure that the project is able to meet the technical objectives. However, key areas where additional resource and/or reallocation of effort is needed is communication, engagement, and training. The operational plan, in section 3.4, describes development and outreach activities, but these are aimed at outreach outside of LTER. The cyberinfrastructure development is a $6M project in its own right, and the social issues of communication and engagement within LTER, engagement of users of the NIS, and training of those users cannot be underestimated. There is significant literature in the area of technology adoption, and LTER can leverage existing relationships previously and being developed with social scientists (e.g., Helena Karasti, Steven Jackson, and Karen Baker) to help achieve these types of goals. The active management of the intra-LTER cyberinfrastructure communication plan should be a part of an LNO staff member's job description.

The project focuses on what needs to be the key goal of the cyberinfrastructure development: making LTER data more accessible to LTER scientists to achieve the science goals of the Decadal Plan for LTER. However, those science goals for cross-site synthesis will need to incorporate data that originate outside of LTER, and the cyberinfrastructure plan needs to more explicitly represent how those data will be discovered and used, particularly in the context of derived datasets that are then folded back into the LTER NIS.

3) Is the time line reasonable to accomplish the project as outlined?

(Discussion Lead: Mike Piasecki) The group felt that the suggested time line (Gantt chart and milestones in text) is quite tight, has no contingency, and leaves little room for breakdown. This

may be a necessary presentation in order to satisfy expectations of what should be done vis-à-vis the received level of funding, but the scope and complexity of the proposed work in all likelihood will encounter set-backs and difficulties slicing into the time line. There is currently no provision for delays that may put an unreasonable burden on the management team to deliver all that was promised. In addition, there is no risk assessment that would outline what could go wrong nor is there any room for action items that the group felt would need to be addressed, e.g. community and site involvement. In summary; in light of the given task of assessing the implementation plan, there is the concern of the timeline being too tight. The group recommends that the timeline be "unstressed" to leave some contingency room for delays. A potential approach is to identify the priority in which implementation areas might be descoped, as a response to the materialization of schedule risks.

4) Is the time line reasonably coordinated to arrive at the desire project goal?
(Discussion Lead: Mike Piasecki) The Gantt chart is quite helpful in assessing this and the group felt that the suggested coordination outline is reasonable. However, the group also felt that a more detailed description that better delineates the dependencies among the various development tracks and their respective tasks, subtasks, and milestones may help seeing critical points in the coordination. In this regard it is important to note the dependencies between broader community involvement (be it defining end products, workflow scenarios or the portal functionality) and NIS development milestones that must be worked into the timeline. Because of the involvement of individuals outside the LNO there is a certain risk in this not happening in a timely fashion thus stressing the timeline. The group's recommendation is to i) add the community involvement into the time line and ii) to better break down tasks and subtasks with the aim of working out dependency points.

5) Are there unidentified risks that lie in the critical path?
(Discussion Lead: Mike Piasecki) The panel felt that the LNO team seems to be aware of some of the risks in this project; however, little was explicitly written down in the plan. As such none of the risks possible have really been identified; hence "unidentified" risks concern really all risks that are there. The panel felt that there are a number of risks that are of particular concern and are identified below, even though we recognize that LNO identified aspects of these risks in their presentation.

- The need to successfully involve the community and sites in the development of workflows, use cases, and expectations as well as GUI functionality.
- The tight timeline (see answers 3 and 4), which leaves little room for contingency and relies on a seamless flow of products and achievements of milestones.
- The relatively small development team targeted for achieving all of the proposed work. There is a risk of hiring in adequate personnel or a high turnover rate because of losses to the workforce. The turnover risk may be particularly acute at the end of the project, depending on LNO's ability to fund positions when this stimulus funding ends.
- Hardware and software failures that render the time schedule obsolete, as well as delays in acquiring the necessary hardware and software.

There are several risks that relate to the interdependencies of this work with other projects. Given the limited resources of this effort, the NIS will need to leverage in efforts from other efforts, such as DataONE and SonNET. Collaboration with these external projects has benefits in terms of more effective use of federal investments, but does create schedule risks. LNO will

need to be strategic and disciplined in determining the appropriate level and scope of collaboration with these related projects.

Given the nature of the funding profile for this project, with allocated funding to accomplish NIS implementation followed by a transition to operational funding, there are particular risks associated with managing that transition to operations, apart from the personnel retention risk identified above. The operational plan generally provides sufficient detail for the assessment of the project, but LNO will need to internally manage a higher level of detail, particularly in terms of the implementation velocity. It will be important to maintain an up-to-date understanding of the level of functionality that will be achieved in this project. Assuming that the currently anticipated funding profile is maintained, the latter part of the project will need to have a particular emphasis on documentation and hardening of implemented functionality so that it can be sustained within an operational budget framework. Management of scope and this regular internal review of the current project timeline is an important element of ensuring that appropriate time and resources are available for project close out activities.

Further discussion on risk and risk management is provided on page 8.

6) Are there standard technologies that are not mentioned, but that we should be aware of? (Discussion Lead: Robert Sandusky) The operational plan has "Web presence improvements for collaboration and presentation" as one major task. Web-based interfaces play an important role in ongoing communications between the LNO and the LTER sites. The review panel recommends that Web presence improvements commit to adherence to prevailing standards, specifically XHTML, CSS, and accessibility / universal design based upon the World Wide Web Consortium's Web Content Accessibility Guidelines (WCAG) 2.0, and OGC standards such as those published for web services (WCS, WFS, WPS, and others). The panel also recommends that the LNO implement these improvements using a standards-compliant content management system to ease maintenance and maximize information re-use. Collectively, these recommendations will result in an LTER / LNO Web presence that is functional across platforms, including mobile devices; is amenable, through standards compliance, to automated processing and will therefore be easier to manage; and is usable and adaptable to people with a wide range of physical and cognitive abilities. Adoption of a content management system also has potential to provide operational efficiencies to the individual LTER sites.

The operational plan has "Maintain network databases" as another major task. The review panel recommends that the LNO consider adopting the Open Archives Initiative Object Reuse and Exchange (OAI-ORE) as a means of linking entries in the bibliographic database, typically published papers, to datasets and other materials, such as related project proposals and reports, that can provide enriched context to scientists interested in re-using LTER data and derived datasets[1]. The LNO should also evaluate the utility of digital object management tools, such as Fedora Commons, in the overall architecture for the NIS, particularly in the context of emergent research linking such tools to content management systems (such as the Islandora project which specifically links Fedora Commons to Drupal).

[1] "The ORE Model introduces the Resource Map (ReM) that makes it possible to associate an identity with aggregations of resources and make assertions about their structure and semantics." (http://en.wikipedia.org/wiki/Open_Archives_Initiative_Object_Reuse_and_Exchange)

Many related metadata technologies are pertinent to sharing data across the LTER Network and with its partners that may augment EML. We encourage the LNO to consider supporting technologies in the Network Information System that focus on sharing sensor-derived, near real-time data streams (for example, sensorML and the related OGC services). With the ability to ingest or distribute sensor-derived data sources, the LTER Network will be poised to interoperate with related networks such as NEON who are also adopting sensor-based technologies.

Particularly with respect to ClimDB, the LNO should evaluate support for the Climate Forecasting convention for the netCDF file convention (netCDF-CF; http://cf-pcmdi.llnl.gov/). Synthesis efforts involving climate data, in particular, will be enabled by providing support for LTER researchers who wish to use ClimDB data currently stored in formats other than the netCDF-CF convention. Likewise, LTER data will be more useful to the broader climate research community if LTER climate data is available in the CF convention. It may be necessary to propose extensions to the CF convention to accommodate LTER data, and this issue should be investigated as early as possible, particularly to enable support for AR5. Similarly, the LNO should consider support for Data Access Protocol (DAP, such as that implemented in OPeNDAP, THREDDS, and other DAP-compliant servers). Both the climate and oceanographic communities are leveraging the power of DAP as a transfer standard, and related projects such as DataONE are actively incorporating the format into their technology stacks, which may help ease adoption within the NIS.

done

has this been done?

Mark

Mark

In addition to these above-mentioned standards, the LNO should consider the appropriate level of engagement for this project with particular related efforts:

- The Earth System Grid (ESG) project, led by Dean Williams at Lawrence Livermore National Laboratory. ESG is based in the climate modeling community and is actively engaged in a range of efforts for integration of model data with field observations and remote sensing data. Collaboration on standards development and implementations of particular protocols is likely to drive both LTER and ESG science objectives and make LTER data more relevant for upcoming IPCC assessments.
- The Federation of Earth Science Information Partners (ESIP) has a number of efforts relating to information standards, with very active work in air quality, water quality, education and outreach, socioeconomic aspects of climate change, and decision support. ESIP is a cross-agency effort with a range of data providers and users dealing with practical aspects of ecoinformatics standards.
- The NASA Earth Science Data and Information System (ESDIS) project, which encompasses the twelve NASA Earth Science data centers (colloquially referred to as the DAACs). The existing collaboration with the ORNL DAAC supports a broader collaboration with ESDIS, and more specific interactions with other data centers may be useful for the LTER science goals. A particular collaboration to consider within ESDIS is the SocioEconomic Data and Analysis Center (SEDAC) housed at the Center for International Earth Science Information Network (CIESIN) at Columbia University. Given the number of LTER sites in the cryosphere, active engagement with the National Snow and Ice Data Center (NSIDC; also a part of ESDIS) may be appropriate.
- Efforts led by the National Climatic Data Center (NCDC) in collaboration with ORNL for integration of remote sensing data with observational data and model data. This is

how about this?

aimed primarily at the calibration and validation of remote sensing data, but the tools are likely to support cross-site integration efforts, particularly in collaboration with those LTER sites that are also Ameriflux sites.

7) Will the steps outlined, if completed successfully, produce a system that achieves the goals stated in the LNO Operational Plan?
(Discussion Lead: Terry Benzel) The operational plan describes a series of integrated tasks in terms of the activities, outcomes, operational steps, milestones, processes for reviewing progress, benefits, and mechanisms for feedback. To a very large extent, this plan is sound and it is expected that the specific steps in the operational plan will result in the development and operation of the CI for NIS. Comments on certain aspects of the plan are discussed in response to the answers above.

The vision and goals for the LTER Network Information Plan are ambitious and can lead to transformative changes in data acquisition, data management and data synthesis for the community. However in order for this vision to be achieved, full community engagement must occur from the beginning of the project. Community involvement must embrace LTER Site Managers, Information Managers, PI's, domain scientists, and the students and postdocs who will be the next generation of researchers. The panel recommends that the LNO evaluate methods and approaches to increasing community engagement, training and communication as a top priority. The operational plan repeatedly includes steps for developing use-case scenarios, requests, working groups, stakeholders and depends on a rich source of feedback from the community. The panel raised some questions as to how well this process will occur and whether a large enough segment of the community will be engaged.

It is strongly recommended that there be a focus on outreach to the sites in order to create a community of active participants. A number of suggestions were discussed during the review, including approaches to getting graduate students in domain science involved at each site, training, tutorials, the use of web presentations, videos and mock ups. These types of activities are needed to ensure that the NIS developments will properly support the LTER science goals, demonstrate capabilities to the stakeholders, and ensure the adoption of the created technologies. As noted in the summary section, the type of outreach envisioned here is different from that of the communication and outreach described in 3.4.2, which appears to be primarily outwardly focused and of a PR nature. The recommendations here would be included in and could benefit from the Strategic Communication Plan described in 3.4.1.

done

In the final analysis, however it appears that a significant limiting factor for community engagement with the LTER sites is an IT resource issue. The individual sites need investment in staff and infrastructure to make NIS participation a seamless part of their everyday activities.

Additional Discussion
This section provides additional information for LTER to consider in proceeding forward with the CI plan. The discussion lead noted for each section is the primary author of that section, drafting a response that represents both that individual's particular expertise and the consensus of the review panel.

Risk Assessment
(Discussion Lead: Michael Piasecki) There is little formal risk assessment in narrative form within the operational plan and also little concerning mitigation strategies for these risks. Formal Risk Assessment however, is an essential component of any operational plan and as such should be included in the write up. The panel encourages the LNO team to include a separate section in their operational plan for a formal risk assessment. The panel does not believe that all risks need to be assessed in detail, but perhaps the most important 5 or 6 risk items that are classified as severe, medium, or light and how a mitigation plan looks like for each of these risks, including resources such as contingency funds or time allotments. In addition to the severity, the risk assessment should also evaluate the likelihood of the risk materializing, the ease by which risk materialization can be recognized in time for corrective action, and the extent to which corrective actions are within LNO control. Such analysis is intended to divert attention from dramatic, but highly unlikely and/or uncorrectable risks. Instead, the focus should be on metrics which indicate the likely materialization of risk and on the plans and processes for reducing the likelihood of such risks materializing, detecting that materialization early, and mitigating the impact of materialized risk.

The panel realizes that contingency funds have not been set aside and, as such, may be hard to identify within the existing budget, however, the LNO team should try to identify existing budget items that could potentially be turned over to become contingency funds. The same is true for the timeline where failures to reach certain milestones could be compensated for by re-arranging the task sequence so as not to cause time delays that cannot be recovered elsewhere.

While the panel realizes that the addition of a risk assessment portion adds extra work, the panel also believes that it will strengthen the operational plan. It will at the very least suggest to the funding agency and potential future site visit reviewers that the LNO team is aware of risks, that they have assessed them and that they have thought about what to do about them in case they emerge. For the sake of time and effort savings the LNO team may just summarize them in a table and provide some short bullets suggesting mitigation steps, but it should be there.

Security
(Discussion Lead: Randy Butler) It is the panel's view that the PASTA architecture needs to document security requirements and develop its software security approach. This software security architecture should implement LTER policy to manage risks to the LTER assets including data, identities, and LTER cyberinfrastructure. The 26 LTER sites historically have been organizationally aligned with a regional entity and within each LTER site there are numerous projects and scientists, however they have been loosely federated through the LNO in the last roughly 10 years. The PASTA CI will further support this federation by the development and implementation of a publically accessible LTER data repository. The PASTA CI introduces new security requirements as a result of the strengthening of the federation of LTER sites, along with the introduction of the centralized LTER data repository and the public accessibility of that repository. The PASTA CI along with the data it houses will become a key asset for the LTER project and there is a critical need to develop a security plan to help ensure the protection of this asset.

The operational plan needs to include the development of the PASTA security architecture that identifies PASTA stakeholders and addresses their security requirements. The panel feels that

HAS THIS BEEN DONE
mark

the software security architecture for PASTA should be documented early in the process because it is a fundamental piece that is not easily added at a later point. The PASTA team has already identified a number of key areas of concern including the need for an identity management plan, authorization and access control to data, and resources, and disaster recovery. Additionally the PASTA team has called out the collaboration with the NSF funded CI-LOGON team to help address identity management.

The LNO does today support an LDAP identity management system that provides LTER users with a unique LTER identity. PASTA, however, will serve data to individuals and organizations that are outside of the LTER project so a broader identity management strategy is needed. The PASTA team is encouraged to prioritize their collaborative work with the NSF funded CI-LOGON team to develop a coherent PASTA security architecture that will both protect the LTER assets, including data and CI, as well as providing safe secure policy enforcement mechanisms to the data.

Community Engagement
(Discussion Lead: Randy Butler) Transforming community engagement also includes a cultural shift. The LTER CI will provide an increasingly sophisticated and capable data analysis CI platform, which has the potential to catalyze world-class research by providing a shared service for growing categories and numbers of users. This envisioned platform can be maximized through a tightly integrated researcher education and community building activity which should seek to leverage modern concepts of social networking and open collaboration. The success of software development projects aimed at large communities relies at least as much on the social strategy of acceptance as it does on the technology. One of the key lessons learned by recent community CI development projects is that the development needs to be a full partnership between the user community and the CI technologists. Because the point of the project is to advance the user community's ability to conduct research and education, the technologist community must respond to that community's needs. If the users are to adopt the technologies and use them effectively, they must have a primary role in determining their capabilities and functions.

Planning the software development and deployment is important but more important is the sharing of those plans. Sufficient planning is important for two reasons: it helps prevent serious execution problems from developing further down the line, and it helps establish that all-important partnership between technologists and users. It is during the planning process that effective communication and management strategies are developed, and it is also during the planning process that the real issues —what the community wants, what the development team can provide, and what the obstacles are — can begin to be identified. A specific recommendation to the PASTA team is to identify key software component release points and to package those components into a unified PASTA release. This is more of marketing than a technology strategy but it may lead to a richer community engagement and offer PASTA users with a better understanding of the evolving PASTA capabilities. Targeting three or four PASTA releases will assist in documenting to end-users where the development and operational capabilities are, and assist them in making decisions on when it is most appropriate for them to engage.

The PASTA team is strongly encouraged to develop a strong community engagement strategy that includes presenting PASTA and its status at LTER and scientific meetings, providing public demonstrations of PASTA, identifying ways that LTER scientists can partner with the PASTA team to further advance their own needs through some kind of experiment or public demonstrations.

The LTER sites are another key engagement point. As identified earlier there appears to be a limited number of IT staff at the sites that represents a risk for site engagement with respect to site integration, and the enrollment of the site data into PASTA. Adding IT staff to the LTER sites alone is likely not enough, these staff will need to work not only with the existing site IT staff but also very closely with the site scientists to help them in the development of metadata and enrollment of their data into PASTA.

This seems to have been addressed. Mark

A key reference for this effort is a recent book that documents a collection of lessons learned for community CI is Olson, Gary M., Zimmerman, Ann, & Bos, Nathan. (2008). *Scientific Collaboration on the Internet*. MIT Press.

Transparency and Shared Development
(Discussion Lead: Chris Jones) An aspect of community engagement involves the transparency of the technology development process, ease of access and ability to quickly evaluate progress and usefulness by stakeholders. To that end, we suggest that all aspects of the software development process be conducted in an open-source fashion where interested parties are able to evaluate the code, planning diagrams, and application interfaces. Having the ability to readily understand the applicability of the code to related projects and to assess the progress to date is critical in reaching the goal of having an integrated LTER network. Adopting a fully transparent process will help to foster community involvement and will enhance communication within the LTER Network and its partners.

Maximum benefit for the broadest set of scientists will be achieved by building an organization to promote sharing of tools, technology, results, and ideas among researchers contributing to the NIS. To make this community model work, the LNO should explore current and future developments in incentive structures to make sharing "profitable" for the user community. Likewise, development of the NIS should pay keen attention to extensibility where researchers with particular domain expertise have a mechanism to integrate their workflows into the larger framework. Success of this community model can dramatically change the landscape from one in which the NIS infrastructure developers are the sole suppliers of all components to one of sharing and mutual community benefit.

This sharing model should go beyond the re-use of technology and methods developed by others; it should be a fundamental enabler of multi-party, multi-technology analysis.