

Prioritized list of recommendations to CCIWG regarding NACP Data Central

The Data System Task Force (DSTF) of the NACP Science Steering Group (SSG) was tasked at the May 2006 SSG meeting with producing a prioritized list of recommendations regarding the plans for NACP Data Central, as an addendum to the Phase 1 recommendations from the DSTF, delivered to the CCIWG in June 2005. The Phase 1 recommendations are available on the Internet: http://www.nacarbon.org/nacp/documents/NACP_Datasystems_recommendations.pdf Based on a series of discussions within the DSTF, the following prioritized recommendations were prepared.

The list below is organized into three categories: (A) high priority functions that are intrinsic to Data Central, (B) high priority functions originally envisioned for Data Central, but that might be met by the thematic data centers (pending selection and identification of thematic centers), and (C) lower priority functions. For each item under the two high-priority categories (A and B), we have included a short statement of explanation / justification.

The Data Task Force focused its deliberations on prioritizing the functions needed for the NACP Data System; the Task Force did not discuss the resources needed for these high priority activities. Should the NACP Science Steering Group or the Carbon Cycle Interagency Working Group require our input on resources, we could provide some specific options along with resource estimates. However, before we could provide options and estimates, we would need to know the data management functions (leveraging) that could be expected from existing and future thematic data centers funded separately by the NACP agencies. In addition, we would like to know the scope and budgetary constraints that would exist for data central.

A. High priority functions intrinsic to Data Central:

1. Coordinate with NACP-related thematic data centers
 - a. Liaise with the NACP project office, NACP investigators, and the thematic data centers to ensure cooperation, mutual understanding of all parties' requirements, and close coordination of data management activities.
2. Identify and/or define metadata standards
 - a. Some very robust metadata standards already exist, and it should be possible to select one of the more widely used standards, with modification or extension as necessary to meet specific needs within NACP. Clearly defined metadata standards will improve the long-term utility of metadata indices as resources to support NACP synthesis and cross-project collaboration.

3. Establish and maintain a metadata index of NACP-relevant data holdings across multiple Thematic Data Centers, including index of data held locally by projects/PIs.
 - a. Given a set of metadata standards, this index of NACP-wide data holdings will play a critical role in supporting synthesis and collaboration across the program. It will allow both NACP investigators and external parties to have a comprehensive view of NACP-relevant data holdings.
4. Coordinate with NACP Office to link the dataset metadata index with the project metadata index generated and maintained by the NACP Office
 - a. This provides critical information for researchers attempting to integrate their own results with other NACP investigations.
5. Provide metadata entry tools for investigators who are not working directly with a thematic data center
 - a. Experience tells us that getting investigators to provide metadata for their data sets is a major hurdle in developing a robust data system. Many such tools exist, but there would be tremendous benefit to selecting these tools carefully, and customizing them as necessary to fit the needs of the NACP data system.

B. High priority functions that might be partially fulfilled by thematic data centers

1. Data versioning system, with subscription and update services
 - a. This is a critical function that anticipates the complex upstream – downstream dependencies between various NACP datasets and projects. Wasted effort can be minimized by keeping researchers informed of new versions of critical data sets.
2. PI-managed access control (following NACP Data Policy)
 - a. This provides a mechanism by which new data sets can be staged into the NACP data system incrementally, with limited release of preliminary uploads and preliminary metadata, followed potentially by revisions and broader access. This provides extra time to work through potential problems with new data sources, as opposed to massive data dumps at the end of a project.
3. Machine access to data holdings (OpenDAP servers)
 - a. This capability is critical, given the very large data volumes anticipated as inputs to and outputs from various NACP projects. Modeling efforts in particular, which will rely on access to large gridded datasets, will benefit from this functionality.
4. Subsetting and GIS data services
 - a. Placing multiple large datasets in a consistent spatial context can be challenging, with many subtle issues that can lead to systematic errors that are difficult to trace. The NACP community would benefit greatly if some core of the most common GIS functions, such as reprojection and subsetting, could be provided through a community toolset.

C. Lower priority functions

1. Automated metadata re-mapping capability
2. Tracking of upstream/downstream data dependencies
3. Housing datasets that don't have another logical home
4. Registry of distributed data services
5. Thematic learning environments (portals)