

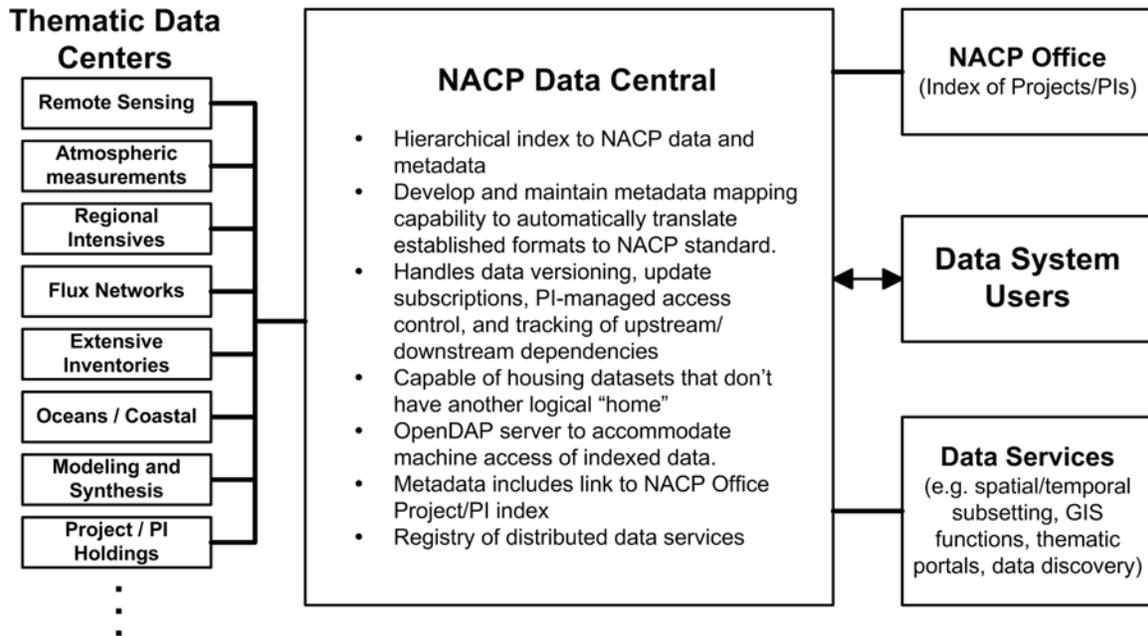
**Phase 1 Recommendations for a  
North American Carbon Program Data System  
NACP Data System Task Force**

The U.S. Carbon Cycle Interagency Working Group (CCIWG) is coordinating the North American Carbon Program (NACP) to measure and understand carbon stocks and the sources and sinks of carbon dioxide (CO<sub>2</sub>), methane (CH<sub>4</sub>), and carbon monoxide (CO) in North America and in adjacent ocean regions. NACP information including the program's science plan and implementation strategy are available on the Internet at <http://www.nacarbon.org/>. Recognizing the need for an integrated NACP data and information management system, in January, 2005 the CCIWG convened a workshop to develop the requirements and architecture for such a system. The workshop report is available on the Internet at <http://www.nacarbon.org/nacp/dm.html>. Following the workshop, the NACP Science Steering Group, in consultation with the CCIWG, appointed a Data System Task Force to further clarify NACP data system requirements and recommend approaches for supporting the initial phase of the program now underway. The following summarizes task force recommendations as of June, 2005.

**Top-Level System Architecture Recommendations**

The diagram below is a schematic representation of the top-level data system architecture recommended by the Data System Task Force (DSTF). Data holdings are clustered by thematic components and served by a distributed network of Thematic Data Centers. Relying on common and/or translatable data formats and metadata standards, this distributed network of data holdings is linked within a comprehensive NACP Data System through the implementation of a centralized metadata indexing system within NACP Data Central. NACP Data Central provides a common access point for retrieving NACP data or metadata.

Data Central provides a common interface for projects to provide data to the NACP community as well as tools for defining product versions, setting access control, and tracking product dependencies. Data Central also provides an index to distributed advanced data services such as geographic information system functions and spatial or temporal sub-setting. Coordination between NACP Data Central and the NACP Office insures that data holdings can be related to relevant project and investigator information.



The following sections present recommended requirements for NACP Data Central followed by recommendations about Thematic Data Centers and Data Services. We identify some existing capabilities as well as certain thematic data components and data service needs that currently are poorly served. A final section lists more general recommendations.

### NACP Data Central Requirements

1. The primary Data Central function is to provide a hierarchical index to NACP data and metadata.
  - 1.1. Metadata copies are held locally within Data Central.
  - 1.2. Most data are held within thematic data centers or with project investigators.
  - 1.3. Data Central also holds datasets without another viable home within the distributed data centers.
  - 1.4. Data Central provides metadata entry tools for investigators not using thematic data center tools.
  - 1.5. An NACP Data System Working Group determines the structure of the indexing hierarchy by selecting a scheme from the several currently in use and potentially modifying that scheme to meet the needs of the NACP data system.
2. Data Central should follow metadata formatting guidelines developed by an NACP Data System Working Group and should:
  - 2.1. Encourage the use of international standards for geographical metadata (e.g., ISO 19115 or the Federal Geographic Data Committee standard) and related standards for other data types. Such metadata standards provide a common terminology and definitions for documenting geospatial data and implementing of such standards within NACP Data Central will facilitate greater access.
  - 2.2. Develop and maintain capability to translate metadata formats from existing data centers into NACP metadata format.

- 2.3. Implement metadata translation tools as new data products are added to the index and as metadata standards change.
3. Data Central should maintain capabilities for tracking data product versions, managing subscriptions to products and updates, implementing investigator managed access control, and tracking of upstream and downstream dependencies.
  - 3.1. On release of new data product versions, indexing within Data Central maintains version control so that older versions, which may have been used by other investigations, remain accessible through the index.
  - 3.2. Data system users can subscribe notices of new datasets added to the index so that they can modify their use of the data as appropriate.
  - 3.3. Data contributors can manage access to the data products, in a manner consistent with the NACP Data Policy. For example, preliminary versions can be made accessible to only a group of collaborators via the index prior to open release.
  - 3.4. Data contributors can add upstream dependency information to their metadata, providing direct links to upstream data products and versions.
4. Data Central should provide technology solutions for machine access to the distributed data holdings (e.g., through OpenDAP servers).
5. Data Central should provide a registry of distributed data services (details below).
6. Metadata on projects and investigators maintained by the NACP Office should be shared with Data Central.
7. Data Central will need to coordinate with the OCCC data management team, including the Volunteer Observing Ship and Repeat Hydrography activities.

### **Thematic Data Centers Recommendations**

1. Thematic Data Centers will hold most NACP datasets populating the Data System.
2. Adopting a distributed system of Thematic Centers for data holdings takes advantage of existing capabilities and thematic expertise, while allowing for the identification and development of NACP thematic components not served by existing centers.
3. The following is a partial list of existing data centers or projects with associations to thematic components of the NACP science strategy:
  - National Center for EROS – remote sensing products, land data products (elevation, land cover, soils),
  - ORNL DAAC – biogeochemical data (field campaign data, regional data, remote sensing products),
  - CDIAC – atmospheric concentration data, eddy covariance flux observations (Ameriflux), county-level fossil fuel emissions,
  - NOAA CMDL – near real-time atmospheric concentration data,
  - WMO World Data Center for Greenhouse Gases – atmospheric concentration data,

- JGOFS Data Center – oceans and coastal data,
  - FIA National Spatial Data Service Center – Extensive forest inventory data,
  - National Resources Inventory – land use and natural resources conditions and trends on U.S. non-Federal lands,
  - State Soil Geographic Database – soils properties and land resource regions,
  - NCDC – surface weather observations, radar summaries,
  - Daymet – grided daily surface weather fields, conterminous U.S.,
  - LANDFIRE – datasets relevant to wildfire science, atlases of U.S. Fire history underway,
  - Land Ecosystem Disturbance Adaptive Processing System (LEDAPS) – reflectance and disturbance data from 1975 – 2000,
  - NCAR Community Data Portal – model data, campaign data, climate-related data sets (biosphere, carbon cycle, atmosphere, land, oceans, etc.),
  - Earth System Grid – access to climate model data on multiple deep storage systems and disks, and
  - Carbon in the Mountains Experiment (CME) – NASA/NSF project that could serve as useful prototype for data management needs of Regional Intensive campaigns.
4. The following major NACP thematic components are not currently served by an existing data center:
- Modeling and Synthesis – Modelers as well as audiences for model output would benefit by having input and output datasets for both forward and inverse modeling, as well as the models themselves and supporting documentation, served in a central location and with consistent formats.
  - Fossil Emission by Source Process – Current emission inventories do not adequately represent the spatial and temporal aspects of various industrial processes of carbon emission sources associated with power generation, mobile sources, heating, and industry.

### **Data Service Recommendations**

1. The following is a partial list of data services that already exist with some functionality at one or more data centers:
- Metadata clearinghouse, data search and discovery (ORNL DAAC, NCAR Community Data Portal),
  - Temporal and spatial subsetting (NC EROS, CDIAC, ORNL DAAC, CMDL, NCDC, Daymet),
  - Reprojection (NC EROS),
  - Web map server (ORNL DAAC),
  - Data compression (NC EROS),
  - Data format conversion, including support for COTS/GIS formats (NC EROS),
  - Summary analyses, trends analyses (CDIAC),
  - Visualization and analysis tools (ORNL DAAC, NCDC, CMDL – e.g. Global View),
  - Real-time weather analysis and data delivery (Unidata), and
  - Grid computing support and common deep storage interface (Earth System Grid).

2. The following data services that are deemed critical by the DSTF, but are not provided at any of the existing data centers:
  - Thematic learning environments (portals) – on-line homes for discussion and learning, organized by thematic topic, and mediated by topical experts. This would include primers or succinct documentation that describes how users used data sets, the problems they had, how they overcame them, etc.,
  - Tracking of product versions and upstream / downstream dependencies, and
  - Comprehensive GIS functionality, allowing visual and quantitative synthesis of data products from across the NACP Data System.

### **Additional Recommendations**

1. **Uncertainty.** The issues related to uncertainty in observations and modeling and how uncertainty propagates through chains of analysis must be dealt with explicitly and prominently in any implementation of an NACP Data System. Investigators should be encouraged to include in their proposals details of how they intend to deal with these issues.
2. **Standards for Models/Modeling.** The NACP Data System must not only foster the optimal use and transparency of measurement data but also the open-source development of modeling tools used in the interpretation of various data sets, leading to improved information products. Model code needs to be developed and maintained for use by others and made available as such by the NACP Data System.
3. **Conventions for Terminology.** To facilitate synthesis and avoid confusion, efforts should be made to standardize data product naming conventions within and across the disciplines that comprise the NACP.
4. **Appropriate Resources.** A general rule-of-thumb from the data management community is that about 10% of a project's funds should be targeted for data management and data services. We provide this as guidance for NACP as a whole, without suggesting that the same level of resource must be dedicated within each funded project or component.

## Appendix A. Task Force Membership

<b>Member</b>	<b>Agency/Organization</b>
Aulenbach, Steve	NCAR
Barr, Alan	Met Service, Canada
Boden, Tom	CDIAC
Cook, Bob*	ORNL-DAAC
Doney, Scott	Woods Hole Oceanographic Institute
Dwyer, John	LP-DAAC
Hollinger, Dave	USDA-FS
Morisette, Jeff	NASA GSFC
Ojima, Dennis	Colorado State University
Pattey, Elizabeth	Agriculture and Agrifoods, Canada
Post, Mac	ORNL
Ramamurthy, Mohan	UNIDATA/UCAR
Reams, Greg	USDA-FS
Tans, Pieter	NOAA-CMDL
Thornton, Peter*	NCAR
Waltman, Sharon	USDA - NRCS
White, Thomas	Pacific Forestry Centre, Canada
*Co-Chairs	