Great Northern Landscape Conservation Cooperative (GNLCC)
Draft Recommended Data Management Plan

Note: This is not a proposal for funding support – this briefing describes ongoing work underwritten by the USFS/FWS Science Support Program

Background

The goal of the Landscape Conservation Cooperative process is to help resource managers address landscape-scale stressors. The issues facing these managers include habitat fragmentation, genetic isolation, spread of invasive species, and water scarcity, as well as how these issues may interact with climate change. Addressing these issues requires access to significant amounts of diverse and well-documented data.

How any Landscape Conservation Cooperative deals with these data issues is a challenge. The areas are vast, requiring collaboration with large numbers of partner agencies and researchers. To make matters more complex, climate assessments can require terabytes of data and produce results not easily integrated with field-collected data.

Objective

This document will provide a data management framework for the Great Northern Landscape Conservation Cooperative (GNLCC). The GNLCC encompasses over 280 million acres, includes parts of 6 states and 2 Canadian provinces, and must address issues that overlap with more than 30 management and research entities. The data management framework will include options for dealing with remotely stored data, datasets integral to resource assessment, and the treatment of model runs.

Data Management in General

All data management plans must provide a framework for data identification, access, acquisition and storage decisions, retrieval, and maintenance. These plans must provide for the ultimate use of the data, satisfying the greatest number of people possible.

The specific approach to data management by any project requires some preliminary steps:

1. *User Identification* – The first step in developing a data management framework is to identify sets of potential data users. Who will ultimately use the data drives every other decision in the data management framework, from what datasets will be addressed, to what type of access routes need to be incorporated in workflows.
2. *Priority Questions* – Once data users are identified, the project must document the priority questions users need to address. A universal catalog of every dataset available in an area of interest would be very hard to work with, and would be extremely expensive to build and maintain. To make any data management framework convenient for users, the datasets addressed should be targeted to the issues the project needs to address. This assessment informs the datasets required, and the extent of project data integration the system must support.
3. *Data Standards and Metadata Requirements* – Accepting all partner datasets without metadata or standards increases the universe of available resources, but can decrease the

quality and the ultimate usefulness of the data. Alternatively, a data quality assessment process which checks all incoming data against defined project data standards and metadata requirements ensures that the data involved in an assessment can be trusted. Use of a data quality assessment process does limit the quantity of data available for an assessment, but is a huge cost savings in the long run. Data quality assessment means: limiting data search to only datasets with complete and compliant metadata; only integrating datasets that meet project timeframe, resolution, and data dictionary requirements; and releasing derived products that are fully documented and conform to all project metadata and presentation standards.

4. *Data Repository and Curation* – Often projects use primary (field data), secondary (processed or derivative data), and/or external products (data from other sources) to generate a result. It is usually not good enough to just view the results; scientists often need to access these primary, secondary, and external products to do their own customizations. One of the decision points of any data management plan is to determine what the scope of the data repository – what data needs to be stored locally versus what will only be a reference to a remotely stored dataset. This decision drives a number of cost options (storage networks, user access rules, and bandwidth needs.
5. *Decision Support and Data Integration* – There are many approaches to decision support, the ultimate goal of any data management system. A number of systems provide ways for the general user to access topical datasets, either through on-line mapping or direct data download. This allows the general user to build custom scenarios and analyses. However, many analyses require significant domain-specific knowledge. Those situations require a specialist to perform the analysis. If there is any likelihood of GIS specialist involvement in the core decision support workflow, the data management system must support data integration options that the specialists are familiar with. These would include cataloging services (e.g., ESRI GeoPortal), networked GIS analysis (enterprise geodatabase and data versioning), on-line product presentation (specialist access to map servers), and model-output consumption and presentation services.

Experience shows that data management for a multi-organization landscape-scale project requires a centralized portal. This allows users to access the quality-assured and pertinent information. The needs of a predetermined group of users help to formulate the questions being asked. This leads to the identification of which data are most important, and how to design the data catalog, data repository, and user interface.

GNLCC Data Management Portal – GIS Specialist-Centric Decision Support

Landscape analysis requires intelligent data association. It is not enough to compile disparate datasets for viewing and distribution; the questions asked by resource specialists and managers normally cannot be answered by cursory view of individual data sets. A coherent data management system must provide pathways for experts to assess, interpret, augment, and associate these data.

This option assumes that data assessment and presentation would be done by a distributed network of authorized GIS analysts. These specialists would work with an integrated data catalog (data search) system and enterprise geodatabase. Select datasets would be augmented or edited, assessed, documented, and served to defined user groups for their specific use. User groups would include resource specialists, other GIS analysts, and managers.

Components

1. *Data Catalog* – The primary portal component would be a cataloging application. This consists of a metadata repository, with search and visualization utilities. Descriptions of datasets are registered with the catalog, allowing users to find these datasets by geographic area, search term, context, source, and/or theme. The catalog application also would have the ability to accept native themes (data that originated with this project and not served by any other source), as well as themes that need to be immediately accessible to modelers.
2. *Enterprise Geodatabase* – The backbone of the catalog service is an enterprise geodatabase. "Enterprise geodatabase" simply means that the data is accessible by multiple simultaneous users, and usually refers to a relational database that is capable of storing and delivering geometry-based information coupled with a file storage area for support data. Native items and submitted themes necessary for analysis would be loaded into (or registered with) this database. Access to the data would be open to a select group of GIS specialists, who would use their GIS client software (e.g., ArcGIS) to augment or customize the datasets and the presentation of those data.
3. *Model Result and Support Data Services* – Not all data needed for an analysis can be captured by a geodatabase. Examples include outputs of climatic, hydrologic, and habitat models. A separate data sharing and distribution mechanism would be developed these data, including capture of the model workflows, inputs, outputs, and assessments. These services could be built using a protocols called Open-source Project for a Network Data Access Protocol (OpenDAP) and Network Common Data Format (NetCDF, used by the climate-modeling community for over 20 years.

Implementation

Establishing the framework consists of the following steps:

- setting up an enterprise geodatabase, one with enough capacity to accommodate the areal extent, resolution, and number of requisite datasets;
- setting up an integrated file store for support data;
- establishing a way for specialists to access the data, along with an access-control strategy;
- implementing data versioning, replication, and backup strategies;
- providing a pathway for GIS specialists to add "intelligence" to the data by
  - correcting and augmenting datasets;
  - adding overlay and presentation information to the theme project;
  - adding metadata to the theme project;
  - defining attribute display, summary, and relation information;
- and providing a network of Thematic Realtime Environmental Distributed Data Services (THREDDS) servers to support NetCDF/OpenDAP outputs from the modeling community.

This approach assumes that the portal also serves as a repository for requisite themes (and not just a catalog of metadata). Normally, the only data stored in a cataloging repository are those themes that are not internet-accessible (no map, download, or catalog server available). This approach changes that model somewhat, accepting themes that the GIS-specialists identify as necessary for problem solving and incorporation into derived products.