Nevada NSF EPSCoR Track 1 Data Management Plan

August 1, 2011

INTRODUCTION

Our data management plan is driven by the overall project goals and aims to ensure that the following are achieved:

- Assure that high quality raw and QA/QC data are produced and are available via the Nevada Climate Change Portal (NCCP);
- Define and apply quality control procedures for all data sets;
- Maintain adequate backup procedures for all data sets;
- Minimize time between data collection, entry, and availability to users;
- Facilitate data access to all project stakeholders, the public, and the global scientific community;
- Ensure data integrity and security;
- Provide metadata for all datasets;
- Perform similar data curation activities and use shared standards and technology across the EPSCoR Tri-State CI Consortium of Idaho, Nevada, and New Mexico.

We recognize the benefits of: (1) freely sharing the data produced during this project with the scientific community, stakeholders, and the public and (2) archiving data in a format that provides maximum interoperability between our collaboration efforts, supports modeling studies initiated by interdisciplinary groups, as well as providing support to the users' community to ease their interpretation and analysis of the data.

This data management plan document describes the procedures and policies for data inventory and data use together with standards for data and meta-data, and privacy. Because the types of data and the needs of users (actual and potential) are so varied from one project component to another, the data management plan includes, where appropriate, component-specific items in addition to general principles.

Definitions

We follow the National Academies (2009) proposed definition of "research data" as:

"Information used in scientific, engineering, and medical research as inputs to generate research conclusions"

Such data includes, but is not limited to:

- Textural information, numeric information, instrumental readouts, images (whether fixed or moving), diagrams, and audio recordings.
- Raw data, processed data, published data, and archived data
- Data generated by experiments, models and simulations, observations of natural phenomena at specific times and locations

- Data gathered specifically for research, as well as information gathered for other purposes used in research
- Data stored on a variety of media

Data inventory

We will provide a complete inventory of all the datasets generated by the Nevada EPSCoR climate change project and available via the NCCP. Datasets may include, but are not limited to, time series data (e.g. from transect weather stations and other sensors); synoptic or survey sampled data; GIS coverages or other spatial datasets; time varying regular or irregular gridded data (e.g. climate model output); and social science data (e.g. responses to surveys). The full data inventory will be available via the Nevada Climate Change Portal, together with information on how to access these data.

General Data and metadata standards

We will utilize existing resources and standards wherever possible, including those developed by CUASHI – Hydrologic Information System (HIS) Observations and Data Model; and the Federal Geographic Data Commission (FGDC) Metadata standard for Digital Geospatial Metadata (CSDGM).

The CI working group established by the Tri-State Consortium (NV, ID, NM) has already established guidelines for common metadata standards. These include: (1) the FGDC Content Standard for Digital Geospatial Metadata (CSDGM), Version 2 (FGDC STD:001.1998) which is already in use by the geospatial clearinghouses in ID and NM; and (2) the ISO 19115:2003 geospatial metadata standard. Given the current use of FGDC metadata in a variety of activities within ID and NM, it was decided that metadata conformant with the FGDC standard would still be produced, although with an understanding that all three states will migrate new metadata production to the ISO19115 standard. This strategy will ultimately entail the conversion of FGDC metadata to the ISO standard, but a variety of tools exist for performing this conversion. As opportunities arise for the development of new metadata management and generation systems over the course of the Tri-state collaboration, the focus will be on those that are compatible with the ISO standard.

The Tri-State CI working group identified two general web service models for use by the states in the development of custom services. These are, in order of preference based upon simplicity of implementation, the Representation State Transformation (REST) web service interface model, the Simple Objects Access Protocol (SOAP) standard, and related web-service standards that have come out of the World Wide Web Consortium (W3C). The CI-Team decided that when selecting standards to implement for particular products and services, the preference would be for existing open standards that are optimized for the specific data (e.g. OGC standards for geospatial data, FGDC/ISO standards for geospatial metadata). The REST and SOAP service models will be adopted when other, more specific standards are not usable.

Three national networks were identified that will be used for tri-state data product development standards: CUAHSI, USGS CLICK, and GEON/OpenTopography. The CUAHSI Open Data Model (ODM) and service interfaces will be used when water-related data are managed and exchanged .

Component-specific Data Standards

Climate Modeling

The Climate Modeling Component will use the following standards for their primary datasets:

NetCDF Files. NetCDF is a very widely used file format that has been specifically designed for storing scientific data, including model output. It is a flexible, self-describing format that bundles data together with any relevant metadata. It can be used to store virtually any kind of data including spatial grids, time-indexed grid stacks, time series, time-indexed profiles (e.g. soil moisture vs. depth) and time-indexed "cube series". When appropriate, data is stored efficiently and compactly in binary form with transparent byte order conversion as needed. Due to the flexibility of the format, several standards have emerged that specify protocols for handling issues such as unstructured grids, names of physical quantities, inclusion of units, and so on. Among these is the so-called Climate and Forecast (CF) convention conformant, see: http://cf-pcmdi.llnl.gov/ or http://www.unidata.ucar.edu/software/netcdf/workshops/2009/cf/index.html. The netCDF format has become so widely used within the scientific community that it is supported by virtually all analysis and visualization packages, including MatLab, IDL, Python, and NCAR Common Language (NCL; http://www.ncl.ucar.edu/Applications/wrf.shtml). In addition, APIs have been developed for virtually every programming language to simplify reading and writing data to netCDF files. There are many other supporting tools for netCDF such as NetCDF Markup Language (NcML) and an online tool for checking whether a netCDF file complies with the CF conventions.

CSV files and Multi-column Text Files: Some simple tabular data will be provided using ASCII column text files. These are human-readable and in an appropriate format that will be implemented for small size post-processed (synthesized) data.

Ecology and Water Components

Sensor data transmitted from the instrumented transects, data collected through manual observations [e.g. field notebooks], metadata, asset management information, maintenance logs) will be stored on the Nevada Climate Change Portal servers. Raw binary and ASCII data (including erroneous data) will be incorporated into a geospatially capable SQL database, which will be registered online with the Knowledge Network for Biocomplexity (http://knb.ecoinformatics.org/index.jsp) and conform to the ecological metadata standards and documentation structure of the Ecological Metadata Language (EML).

Education Component

For each of the datasets generated the following information will be provided: title, data collection date, author(s), abstract/study description, keywords, temporal coverage, geographic coverage, contact, methods of production, types of surveys, codebook/data dictionary, background material (e.g. sample survey questionnaires, report), and html download link if applicable.

Policy and Outreach Component

This component will generate at least 8 primary datasets, comprising the results of surveys of the attitudes of different groups to climate change. Groups surveyed include farmers and

ranchers; Native Americans; Tribal environmental Managers; Public Agencies; NSHE Students; Business leaders; Water Purveyors; and Energy Purveyors. Data will include survey responses (without personal information), location information. Metadata will include the survey design.

Data Use, Privacy and Sharing Standards

All project members recognize the benefits of freely sharing data with the scientific community and interested stakeholders. Users of project data will be asked to acknowledge use of datasets in publications and reports, and to provide the project with copies of these documents. Data use and sharing will also be subject to Nevada System of Higher Education and the National Science Foundation policies pertaining to intellectual property, record retention, and data management.

Each component will designate a contact person for questions or to report problems related to datasets.

Climate Modeling Component

All of the Climate Models, model data, and observations stored and distributed by the Climate Modeling Component will be available under open-source. Users will be asked to acknowledge the use of the data as follows:

> *Acknowledge your use of the data in any documents or publications using these data. We would also appreciate receiving a copy of the relevant publications and white papers. Thank you!*
>
> *"Regional Climate Modeling simulated output provided by the DRI Climate Modeling Group, Reno, Nevada, USA, and distributed through #Web Portal#."*

We will also request some information from users with regard to their research using DRI-RCM and DRI GCM simulated data: correspondence and affiliation; main research goals; methodology; and funding institution. This information is only for internal use and will help us to guide the type of products and data needs from the user's community. Further, we will create an email list to redistribute new versions of the data and additional products.

Ecology and Water Components

All sensor-generated data (e.g., raw data directly from the data loggers, QA/QC checked data, and Web camera images) will be provided to end-users (e.g., researchers stakeholders, teachers, and the general public), in a timely manner, as close to real time as possible, via the Nevada Climate Change Portal. Access will be facilitated by a graphical user interface that will allow users to see most, but not all, data in near real time and download data. The Data Portal will allow all users to select the type of data they need (e.g., binary, tab- or space-delimited, ASCII, XHTML, EML, .JPG, or XML formats). An example of data output might include scalar values (e.g., wind speed, precipitation, soil moisture), timestamp, geospatial region, originating sensor, transect site, or other relevant information needed by the individual downloading the data. Data will be provided without assurance of data quality. Users of data from the project must obtain written permission from either one of the Co-PIs or from one of the Leads for the Water Resources and Ecological Change components before publishing the data. We anticipate, however, that our open data access data policy will enhance publication of data generated by the proposed study by actively

encouraging scientists, students, and even stakeholders to collaborate in writing peer-reviewed papers together with project researchers.

Education Component

Data will remain within the project for two years or until publication (whichever is first). Two types of data will be provided: public data and restricted data. The Professor Survey Questionnaire and Climate Education Conference Outcome will be uploaded to the Data Portal and be available for the public.

For the Professor Survey raw and processed data, and Climate Change Course Inventory, users of these data will be asked to obtain written permission from the Education Component Lead: Dr. Michael Nussbaum, Email: nussbaum@unlv.nevada.edu, Phone: (702) 895-2665, Fax: (702) 895-1658."

The Professor Survey and Climate Change Course Inventory contain personal contact, demographic, and individual perception information. The survey respondents have been informed that this information would be kept confidential when they filled out the survey. The sharing and use of this type of data will therefore be restricted.

Policy and Outreach Component

Data will remain within the project for two years or until publication (whichever is first).

Two types of data will be deposited with the Data Portal: public data and restricted data, which will become public either through publication or after two years. All the survey questionnaires with released reports and theses or dissertations/articles will be uploaded to the Data Portal and be available for the public. Users of raw and processed survey data must obtain written permission from either one of the PI/Co-PIs or from the Policy, Decision Making and Outreach Component Lead.

The surveys contain demographic and individual perception information. The survey respondents have been informed that their information would be kept confidential before they fill out the survey, and all identifying information has been removed. Any publication or use of the material that attempts to identify specific individuals is not permitted.

Data Management Life Cycle

As far as is possible, the final storage location for all data from the transects (e.g., sensor data transmitted through the data network, data collected through manual observations [field notebooks], metadata, asset management information, maintenance logs) will be the Nevada Climate Change Portal servers located at the University of Nevada, Reno.

Backup storage and archiving of sensor data from the instrumented transects will also be facilitated by the Western Regional Climate Center, located at DRI, Reno.

The large volumes (hundreds of TB) generated by climate models will require special attention. The Climate Modeling group has a high-performance computing cluster (HPC) named "GridLogin" that is available to this project. This system has 380 TB of RAID hard drive storage, 100 TB of which are automatically backed up on a regular basis to LTO4 tapes. Each tape holds up to 2 GB with compression. Some storage space will be reserved to archive and distribute relevant observed and simulated data. Climate model output needed storage space ranges up to several hundreds of TB, which is larger than the storage space reserved in the Nevada Climate Change Portal servers. Therefore, our strategy will use the

data portal for distribution of model output, with archived the data in GRidLogin storage space.

Survey data from the Education and Policy and Outreach components will be archived via the Nevada Climate Change Portal.

Useful URL's

CUASHI data management guidelines

http://www.cuahsi.org/his-dmp.html

Reference

National Academies of Science, Engineering, and Medicine, 2009. *Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age*, National Academies Press, Washington DC.