

Example Data Management Plan: NSF General

Product of a DataONE Best Practices Workshop.

Background Information (not part of a DMP)

Title:

Rio Grande Basin Hydrologic Geodatabase Compendium.

Purpose of project:

The U.S. Geological Survey (USGS), in cooperation with the New Mexico Interstate Stream Commission, developed a geodatabase compendium (hereinafter referred to as the geodatabase) of available water-resources data for the reach of the Rio Grande from Rio Arriba-Sandoval County line, New Mexico, to Presidio, Texas. Since 1889, a wealth of water-resources data has been collected for a variety of purposes. Collecting agencies, researchers, and organizations have included the USGS, Bureau of Reclamation, International Boundary and Water Commission, state agencies, irrigation districts, municipal water utilities, universities, and other entities. About 1,750 data records were recently (2010) evaluated to enhance their usability by compiling them into a single geospatial relational database (geodatabase). All data available, including water quality, water level, and discharge data (both instantaneous and daily) from January 1, 1889, through December 17, 2009, were compiled for the study area. Geodatabase tables include daily data values, site locations and information, sample event information, and parameters, as well as data sources and collecting agencies. The end products of this effort are a comprehensive water-resources geodatabase that enables the visualization of primary sampling sites for surface discharges, groundwater elevations, and water-quality and associated data for the study area. In addition, repeatable data processing scripts, Structured Query Language queries for loading prepared data sources, and a detailed process for refreshing all data in the compendium have been developed.

For more information on the project, access the report here: <http://pubs.usgs.gov/of/2010/1331/>

Data Management Plan

Rio Grande Basin Hydrologic Geodatabase Compendium.

1. Types of Data Produced

This project will result in the production of a relational spatially-enabled database integrating all known surface water, ground water, and water quality data for the middle Rio Grande basin study area. Additionally, Visual Basic for Applications (VBA) code and Structured Query Language (SQL) code are products of the project. All updateable datasets are acquired from the original data source (for example, EPA websites). Updateable data sources are acquired at specified intervals – quarterly, or as needed. As new static data sources are discovered, they will be integrated into the proposed compendium. Data will be processed using dataset-specific VBA programs. Program file comment headers will be included in the code to explain required input variables, purpose of the program, and requirements needed by the user. Code will be annotated to promote code readability.

2. Data and Metadata Standards

Microsoft Access Database format will be used since it is readily-accessible and it is compatible with ESRI ArcGIS (<http://www.esri.com/software/arcgis/index.html>), a Geographic Information System software package used by the stakeholders. Naming conventions will be consistent – no spaces will be used in table names or field names. The file naming convention will consist of the data *source_data* type format for raw data files. Data reporting functionality will be built into the VBA processing programs to provide output in .txt file format for number of records per source when updateable data sources are refreshed.

Every effort will be made to go back to the authoritative source for an identified dataset. Quality control of the database will be performed using SQL statements that capitalize on the database structure to ensure relational database integrity. Appropriate primary keys will be assigned to manage possible data duplicates. Potential duplicate site IDs, will be handled through automated procedures and the creation of alternate ID tables.

A data dictionary will be created that defines the table definition, table fields, and table field data types. An entity-relationship diagram will be created that defines the relational structure of the database. A metadata record will be produced using the FGDC standard that describes the entire geodatabase.

The FGDC standard was chosen due to required Federal government standards.

3. Policies for Access and Sharing

The data are public and will be obtainable thru the New Mexico Interstate Stream Commission (NMISC). Users of the data will primarily be water resource managers in the Rio Grande Basin. USGS publications will be released describing the methods and data sources and can be used as documentation for the data and to cite the data.

4. Policies for Re-use, Distribution

Access to databases and associated software tools generated under the project will be available for educational, research and non-profit purposes. Such access will be provided using web-based applications, as appropriate.

Materials generated under the project will be disseminated in accordance with University/Participating institutional and NSF policies. Depending on such policies, materials may be

transferred to others under the terms of a material transfer agreement.

The data files have a suggested citation, which will be described in the metadata in addition to the USGS publications.

5. Plans for Archiving and Preservation

All original raw data files and data source processing programs will be versioned over time and maintained in a date-stamped file structure with text files documenting the provenance. The database will be preserved in perpetuity, housed initially at the New Mexico Interstate Stream Commission Central Office in addition to an off-site copy maintained at an NMISC field office and mirrored at the Consortium of Universities for the Advancement of Hydrologic Science (CUAHSI). We will also identify appropriate archiving institutions that might serve as a mirror repository. A data policy and stewardship plan will be established. In addition to archiving, each database table will be exported to a delimited text format to ensure accessibility of the data by other software programs. The data manager at the NMISC will be responsible for the management of long-term storage and archived data.