Plan Overview

A Data Management Plan created using DMPTool

Title: Boulder Creek Critical Zone Observatory Data Management Plan

Creator: Jeri Fey

Affiliation: University of Colorado Boulder (CU Boulder) (colorado.edu)

Funder: National Science Foundation (nsf.gov)

Funding opportunity number: 16058

Template: NSF-EAR: Earth Sciences

Last modified: 08-13-2015

Boulder Creek Critical Zone Observatory Data Management Plan

Types of data

Preservation of all data, samples, physical collections and other supporting materials needed for long-term earth science research and education is required of all EAR-supported researchers.

The Boulder Creek Critical Zone Observatory (CZO) focuses on research in the Boulder Creek watershed. This encompasses Green Lakes Valley, Gordon Gulch, and Betasso locations covering 1158 km2 at 1480-4120m of elevation. There are two groups of data collected. The first is ongoing data collection that starts in 2008 and is comprised of manual sample collection, manual measurements and data loggers. The second is completed project data.

Ongoing data collections are comprised of:

- Meteorological data - which includes historical and live data comprised of air temperature, humidity, wind speed and direction.
- Snow Depth data - from Judd snow depth sensors and manual in situ measurements
- Electrical Conductivity data - from soil moisture and temperature data loggers
- Surface water Chemistry data – from lab analyzed samples
- Snow Pit data – manually collected density and stratigraphy
- Stream Flow/Discharge data - both manual observations as well as data loggers
- Time Lapse Camera data
- Well water level data – both manual and data loggers

Completed project data collections are comprised of:

- Diatoms
- Dissolved Organic Matter from lysimeter samples
- Snow Survey of the water shed
- Soil Geochemistry
- Soil Microbes
- Soil Respiration
- Tree growth
- LiDAR
- Physiology Geophysical data from Shallow Seismic Refraction and Electrical Resistivity

Graduate students typically collect the completed project data collections with a specific topic in mind, resulting in a published paper. The ongoing data sets are collected by the field manager, lab manager and trained students with the purpose of creating a historical record to be used for any research topic relating to the Earth's critical zone.

Data collected in situ and sample data analyzed in the lab, are subjected to a quality assurance and quality control process before being submitted to the Boulder Creek CZO website for public access.

Data and metadata standards

Data archives must include easily accessible information about data holdings, including quality assessments, supporting ancillary information, and guidance and aids for locating and obtaining data.

All data has been required to be submitted in comma separated value (.csv) format with accompanying meta data file in text (.txt) format. Currently the meta data files are being converted to .csv and .xml files in accordance with ISO-19115 Geographic Metadata standards. The meta data is being modeled from "A Model Information Management System for Ecological Research, Rick C. Ingersoll, Tim R. Seastedt, and Michael Hartman,

BioScience Vol. 47, No. 5 (May, 1997), pp. 310-316" which has been expanded and built upon by the creators of its design since its publication.

Meta data must have the following values:

- Title – for searching capabilities
- Author
- Contact
- Unique Location ID
- Location ID Subset
- Location – either Betasso, Upper or Lower Gordon Gulch, or Green Lakes Valley
- Location Description
- Location UTM – North and South bounding latitudes
- Location UTM – East and West bounding longitudes
- Date range
- Frequency
- Abstract
- Investigator

- Citations
- Keywords
- Methods
- Variables
- Acronyms

If a new data set is submitted then the meta data is used initially to determine which field location, topic and discipline the data should be saved in. If this is an existing data set

that has new data then the log files are updated according to the field manager's notes.

All data sets get their own web page with searchable meta data listed on the page itself as well as available to download in .csv format. Each web page has a link to download the data directly from a .csv file for completed projects. For ongoing data set collections, the data is inserted into an Oracle relational database which can be queried from the website for specific variables and date ranges.

The database and web server are hosted on a server supported and backed up by the data manager and CU's managed services group, which is a division of the Office of Information Technology at the University of Colorado at Boulder.

Policies for access and sharing

It is the responsibility of researchers and organizations to make results, data, derived data products, and collections available to the research community in a timely manner and at a reasonable cost. In the interest of full and open access, data should be provided at the lowest possible cost to researchers and educators. This cost should, as a first principle, be no more than the marginal cost of filling a specific user request. Data may be made available for secondary use through submission to a national data center, publication in a widely available scientific journal, book or website, through the institutional archives that are standard for a particular discipline (e.g. IRIS for seismological data, UNAVCO for GP data), or through other EAR-specified repositories. Data inventories should be published or entered into a public database periodically and when there is a significant change in type, location or frequency of such observations. Principal Investigators working in coordinated programs may establish (in consultation with other funding agencies and NSF) more stringent data submission procedures.

Every data set is accessible from the http://czo.colorado.edu website and can be searched by title, field location, topic, or discipline. These data sets can also be located using interactive GIS/Map located here: http://czo.colorado.edu/geGIS/gmGIS.shtml

The meta data is what gives the Boulder Creek CZO its searching power. This searching capability is also ported to the national CZO site where all of the Boulder Creek CZO data sets are available in addition to data sets from nine other CZOs. Each CZO uses the same meta data formatting in order to be searchable from the national level here http://search.criticalzone.org/.

Each web page provides a description, keywords and citation that can be used for searching or reporting from the data set. There is a data use policy posted on every data set page that explains how to use or re-use this data. Which adheres to NSF's policy on dissemination.

Data Use Policy:

1. ***Use our data freely**. All CZO Data Products* except those labelled Private** are released to the public and may be freely copied, distributed, edited, remixed, and built upon under the condition that you give acknowledgement as described below. Non-CZO data products — like those produced by USGS or NOAA — have their own use policies, which should be followed.*
2. ***Give proper acknowledgement**. Publications, models and data products that make use of these datasets must include proper acknowledgement, including citing datasets in a similar way to citing a journal article (i.e. author, title, year of publication, name of CZO "publisher", edition or version, and URL or DOI access information. Seehttp://www.datacite.org/whycitedata).*
3. ***Let us know how you will use the data**. The dataset creators would appreciate hearing of any plans to use the dataset. Consider consultation or collaboration with dataset creators.*

**CZO Data Products. Defined as a data collected with any monetary or logistical support from a CZO.*

***Private. Most private data will be released to the public within 1-2 years, with some exceptionally challenging datasets up to 4 years. To inquire about potential earlier use, please contact us.*

Policies and provisions for re-use, re-distribution

For those programs in which selected principle investigators have initial periods of exclusive data use, data should be made openly available as soon as possible, but no later than two (2) years after the data were collected. This period may be extended under exceptional circumstances, but only by agreement between the Principal Investigator and the National Science Foundation. For continuing observations or for long-term (multi-year) projects, data are to be made public annually.

The data for ongoing research data sets are updated monthly for all data loggers, only during the fall and winter for snow data sets, and annually in the summer for time lapse and surface chemistry. Typically the data is collected by the field manager, QA/QC'd and posted online within about 2-3 months for public access.

For completed or original datasets the data owner does have some time to work with the data before it is required to be submitted. Below is the Data Sharing Policy posted on every data set web page. Which adheres to NSF's policy on sharing.

Data Sharing Policy:

1. ***Share data privately within 1 year**. CZO investigators and collaborators agree to provide CZO Data Products* — including data files and metadata for raw, quality controlled and/or derived data — to CZO data managers within one year of collection of samples, in situ or experimental data. By default, data values will be held in a Private CZO Repository**, but metadata will be made public and will provide full attribution to the Dataset Creators†.*
2. ***Release data to public within 2 years.** CZO Dataset Creators will be encouraged after one year to release data for public access. Dataset Creators may choose to publish or release data sooner.*
3. ***Request, in writing, data privacy up to 4 years.** CZO PIs will review short written applications to extend data privacy beyond 2 years and up to 4 years from time of collection. Extensions beyond 3 years should not be the norm, and will be granted only for compelling cases.*
4. ***Consult with creators of private CZO datasets prior to use.** In order to enable the collaborative vision of the CZO program, data in private CZO repositories will be available to other investigators and collaborators within that CZO. Releasing or publishing any derivative of such private data without explicit consent from the dataset creators will be considered a serious scientific ethics violation.*

** CZO Data Products. Defined as data collected with any monetary or logistical support from a CZO. Logistical support includes the use of any CZO sensors, sampling infrastructure, equipment, vehicles, or labor from a supported investigator, student or staff person. CZO Data Products can acknowledge multiple additional sources of support.*

*** Private CZO Repository. Defined as a password-protected directory on each CZO's data server. Files will be accessible by all investigators and collaborators within the given CZO and logins will be maintained by that local CZO's data managers. Although data values will not be accessible by the public or ingested into any central data system (i.e. CUAHSI HIS), metadata will be fully discoverable by the public. This provides the dual benefit of giving attribution and credit to dataset creators and the CZO in general, while maintaining protection of intellectual property while publications are pending.*

† Dataset Creators. Defined as the people who are responsible for designing, collecting, analyzing and providing quality assurance for a dataset. The creators of a dataset are

analogous to the authors of a publication, and datasets should be cited in an analogous manner following the emerging international guidelines described at http://www.datacite.org/whycitedata.

Plans for archiving and preservation of access

Remember - Data may be made available for secondary use through submission to a national data center, publication in a widely available scientific journal, book or website, through the institutional archives that are standard for a particular discipline (e.g. IRIS for seismological data, UNAVCO for GP data), or through other EAR-specified repositories.

For short term archiving purposes this data is backed up nightly and retained for 30 days. However, because of the flat file nature of a UNIX server running an Oracle database and Apache Tomcat web server, the CZO does have full backups created quarterly and saved to external hard drives.

For long term archiving there are a couple of options in place. Currently this is an ongoing funded project which will keep the data available in the near future. This data is also hosted on the National CZO website for the further foreseeable future.

Time series data is formatted so that it can be ingested in the CZO Central Data Portal (Zaslavsky et al., 2011) that forms the center of the CZO Integrated Data Management plan (NSF 1153164 to Aufdenkampe). The National CZO website (http://criticalzone.org) provides the access to the Central Data Portal.

The goal for the Boulder Creek CZO is create and collect meaningful and interesting research of the Earth's critical zone, by making this diverse data, available to the public as soon as it is available. As well as providing access to other CZO's data sets for similar research of the weathered, hydrologically active near surface environment.