Guidance and Requirements for

NCCWSC/CSC Data Management Plans

(Required for NCCWSC and CSC Proposals and Funded Projects)

Prepared by the CSC/NCCWSC Data Management Working Group

Contact: Emily Fort, Data and IT Manager for the National Climate Change and Wildlife Science Center, efort@usgs.gov, 703-648-4082

Version 1.0

Contents

Requirements for NCCWSC/CSC Data Management Plans

Background

The National Climate Change and Wildlife Science Center (NCCWSC) and the DOI Climate Science Centers (CSCs) established a data sharing policy requiring a data management plan for all proposals and funded projects. This document provides guidance on what should be included as part of the data management plan.

The requirement for data management plans has been divided into two stages: 1) the **Proposal Data Management Plan (PDMP)**, which will be submitted as part of the proposal, and 2) the **Research Data Management Plan (RDMP)**, which includes information provided in the PDMP plus additional information about data handling and management throughout the project's life cycle. The PDMP requirements are limited to information needed to understand the proposal from a data perspective. The RDMP includes additional details to ensure that data products are consistently formatted and documented during development, and that data deliverables are provided with maximum utility. The RDMP will help the CSC Data Manager assist the researcher with data management activities and make the project's data and derived products available after project completion. The RDMP should be reviewed and enhanced as the project evolves. The initial version of the RDMP is due within three months of the proposal being funded. Following that, the RDMP should be reviewed and updated quarterly. The PDMP and RDMP include information about data used in proposed projects including: 1) data inputs – existing data collections, 2) data inputs – collections of new field data or model output, and 3) data products and deliverables.

Why Are Data Management Plans Required?

- Assists with reproducibility of research projects because data are well documented.
- Helps ensure data and data products are accessible and available for the long term.
- Consistent with the best practices from many funding agencies such as the National Science Foundation.

Data Management Plan Description and Guidance

For both the PDMP and the RDMP, descriptions of the data are divided into the following four categories:

1. Data Inputs – Existing Collections;
2. Data Inputs – New Collections (e.g., data collected from the field, new model output);
3. Data Inputs – Software or Other Needs (RDMP Only); and
4. Data Outputs – Expected deliverables, datasets, and products.

Information should be provided for each element as appropriate. If a section is not appropriate (i.e., no new data are being collected), that should be noted and no additional information is necessary. If information is not known at the level of detail requested, as much information as possible should be provided. For multiple inputs and outputs, the data description table should be copied as many times as needed so that each table represents one input or one output.

The information provided in the PDMP will be reviewed as part of the proposal process. When proposals are funded, information in the proposal and PDMP will be used to create initial records for the project in the NCCWSC and CSC project tracking tool – ScienceBase. Additionally, a CSC data manager will contact the principal investigator of each funded project to assist with the development of the RDMP.

Template for Proposal Data Management Plan Requirements

Project Title: [Insert Project Title]

Data Inputs – Existing Data Collections

1	[Name of Collection]
Description:	Describe the existing data, model, etc. that will be used. If not known, please provide as much information as possible (e.g., remote sensing, global climate models, etc.).
Restrictions:	Identify any limitations on access or reuse (e.g., sensitive data, restricted data, software with license restrictions, etc.).

Data Inputs – New Data Collection

1	[Provide a brief name to describe new data collection]
Description:	Describe the information to be gathered and the scale (e.g., national, regional, landscape, etc.) of the data.
Data Management Budget:	Describe the proposal budget portion allocated for data management activities for the new data collected.
Protocols:	Identify any standard protocols or methodologies that will be used to collect the data, if available.
Quality Checks:	Specify the procedures for ensuring data quality.
Exclusive Use:	Project data and associated products should be available publically at the end of the project. If a request to limit access for a period of time after project completion is needed, identify the length of time and the reason for the extension. (Request cannot be more than two years.)
Restrictions:	Identify any limitations on access or reuse (e.g., sensitive data, restricted data, software with license restrictions, etc.).
Contact:	Provide a point(s) of contact if questions arise related to the data and associated products (name, email, and phone number).

Data Outputs – Deliverables, Datasets and Products

1	Name of Output
Description:	Describe the deliverable/product.
Data Management Budget:	Describe the proposal budget allocated for data management activities for the new data collected.
Format:	Identify the likely format for the deliverable/product.
Quality Checks:	Specify the procedures for ensuring data quality.
Exclusive Use:	Project data and associated products should be available publically at the end of the project. If a request to limit access for a period of time after project completion is needed, identify the length of time and the reason for the extension. (Requests of more than two years will not be granted.)
Restrictions:	Identify any limitations on access or reuse (e.g., sensitive data, restricted data, software with license restrictions, etc.).
Contact:	Provide a point(s) of contact if questions arise related to the data and associated products (name, email, and phone number).

Template for Research Data Management Plan Requirements (RDMP)

Project Title: [Insert Project Title]

Data Inputs – Existing Collections

1	[Name of Collection]
Description:*	Describe the information that will be used and the nature and scale (e.g., national, regional, landscape, etc.) of the data.
Format:	Identify the formats in which the data are maintained and made available.
Source:	Identify the source for the data.
Data Processing & Scientific Workflows:	Describe any data processing steps or provide a scientific workflow you plan to use to manipulate the data, as appropriate.
Backup & Storage:	Describe the approach for backup and storage of the information associated with the research project.
Volume Estimate:	Estimate the volume of information that will be generated: megabyte (MB), GB, TB, or PB.
Access & Sharing:	Prior to the completion of the project, specify who should have access to project information/products and what type of access (Public, Read, Write, No Access).
Restrictions:*	Identify any limitations on access or reuse (e.g., sensitive data, restricted data, software with license restrictions, etc.). Provide citation or documentation describing limitations if due to policies or legal reasons.
Fees:	Identify any fees associated with acquiring the data.
Citation:	Provide citation for data product.

*Indicates that this information is also collected for the PDMP.

Data Inputs – New Collections

1	[Provide a brief name to describe new data collection]
Description:*	Describe the information that will be used and the nature and scale (e.g., national, regional, landscape, etc.) of the data that will be collected.
Data Management Budget:*	Describe the proposal budget portion allocated for data management activities for the new data collected.
Format:	Identify the formats in which the data will be generated, maintained, and made available.
Data Processing & Scientific Workflows:	Describe data processing steps or provide a scientific workflow you plan to use to manipulate the data, as appropriate.
Protocols:*	Identify any standard protocols or methodologies that will be used to collect the data, if available.
Quality Checks:*	Specify the procedures for ensuring data quality.
Backup & Storage:	Describe the approach for backup and storage of the information associated with the research project.
Metadata:	Identify the metadata standard that will be used to describe the document (FGDC, ISO, EML, etc.)
Volume Estimate:	Estimate the volume of information generated: megabyte (MB), GB, TB, or PB.
Archive Organizations:	Identify the organization that will be responsible for the long-term archive and preservation of the data collected.

1	[Provide a brief name to describe new data collection]
Access & Sharing:	Prior to the completion of the project, specify who should have access to project information/products and what type of access (Public, Read, Write, No Access).
Exclusive Use:*	Project data and associated products should be available publically at the end of the project. If a request to limit access for a period of time after project completion is needed, please identify the length of time and the reason for the extension. (Request cannot be more than two years.)
Restrictions:*	Identify any limitations on access or reuse (e.g., sensitive data, restricted data, software with license restrictions, etc.). Provide citation or documentation describing limitations if due to policies or legal reasons.
Citation:	Specify how the project's data should be cited.
Digital Object Identifier (DOI)/Link:	Provide a digital object identifier (DOI)/link to the project when availably publically.
Contact:*	Provide a point(s) of contact if questions arise related to the data and associated products (name, email, and phone number).

*Indicates that this information is also collected for the PDMP.

Software and Other Needs

1	[Name of Software or Other Need]
Description:*	Describe any software or other needs that are required for the project.
Restrictions:*	Identify any limitations on access or reuse that accompany the software or other needed items.
Fees:*	Identify any fees or other costs associated with acquiring the software or other items.
Source/Link:	Provide a link or a source for the need if available.

Data Outputs (e.g., Project Deliverables or Products)

1	[Name of Output]
Description:*	Describe the data output.
Data Management Budget:*	Describe the proposal budget portion allocated for data management activities for the data output.
Format:*	Identify the formats in which the data will be generated, maintained, and made available.
Data Processing & Scientific Workflows:	Describe data processing steps or provide a scientific workflow you plan to use to manipulate the data, as appropriate.
Quality Checks:*	Specify the procedures for ensuring data quality during the project.
Backup & Storage:	Describe the approach for backup and storage of the information associated with the research project.
Metadata:	Identify the metadata standard that will be used to describe the data and products (FGDC, ISO, EML, etc.)
Volume Estimate:	Estimate the volume of information generated: megabyte (MB), GB, TB, or PB.

Access & Sharing:	Prior to the completion of the project, specify who should have access to project information/products and what type of access (Public, Read, Write, No Access).
Exclusive Use:*	Project data and associated products should be available publically at the end of the project. If a request to limit access for a period of time after project completion is needed, please identify the length of time and the reason for the extension. (Request cannot be more than two years.)
Archive Organizations:	Identify the organization that will be responsible for the long-term archive and preservation of the data collected.
Restrictions:*	Identify any limitations on access or reuse (e.g., sensitive data, restricted data, software with license restrictions, etc.). Provide citation or documentation describing limitations if due to policies or legal reasons.
Citation:	Specify how the project's data should be cited.
Digital Object Identifier (DOI)/Link:	Provide a digital object identifier (DOI)/link to the project when availably publically.
Contact:*	Provide a point(s) of contact if questions arise related to the data and associated products (name, email, and phone number).

*Indicates that this information is also collected for the PDMP.

Appendix A – Proposal Data Management Plan Examples

A-1: North Central Climate Science Center Pilot Project

Data Inputs – Existing Data Collections

1	Daymet2 Climate Dataset
Description:	Daily climate metrics for North America (US, Canada, Mexico)
Restrictions:	None. Distributed by Oak Ridge National Lab.

Data Inputs – New Data Collection

No new collections will be developed.

Data Outputs – Deliverables, Datasets and Products

1	Century model outputs
Description:	Century model ouputs
Data Management Budget:	20% of overall project budget, includes data storage, data organization support, integration, use of ScienceBase project space, backup, replication, and archive
Format:	Csv
Quality Checks:	Validation done via python scripts and VisTrails workflow tool
Exclusive Use:	No exclusions
Restrictions:	No restrictions
Contact:	John Doe, john.doe@university.edu,

2	Dynamic Custom Visualization
Description:	Dynamic charting, graphing, mapping, and reporting on Century model outputs
Data Management Budget:	20% of overall project budget, includes data storage, data organization support, integration, use of ScienceBase subversion repository, backup, replication, and archive
Format:	Java application (.java files)
Quality Checks:	Code review and regression testing
Exclusive Use:	No exclusions
Restrictions:	No restrictions
Contact:	John Doe, john.doe@university.edu,

3	VisTrails Workflow
Description:	Workflow replication and description
Data Management Budget:	20% of overall project budget, includes data storage, data organization support, integration, use of ScienceBase subversion repository, backup, replication, and archive
Format:	.vt files
Quality Checks:	Code review
Exclusive Use:	No exclusions
Restrictions:	No restrictions
Contact:	John Doe, john.doe@university.edu,

A-2: Streamflow Simulations for Hydrological Modeling of Lake Michigan Streams and Impacts on Trout

Data Inputs – Existing Data Collections

1	Daily precipitation and temperature data
Description:	Daily precipitation and temperature data from NCDC (http://www.ncdc.noaa.gov/oa/climate/research/cag3/cag3.html)
Restrictions:	None known

2	NLCD
Description:	Land cover (forest, wetlands, impervious): http://www.mrlc.gov/finddata.php
Restrictions:	None known

3	Water withdrawals and discharges dataset (PCS database)
Description:	Withdrawals and discharges
Restrictions:	None known

4	Streamflow Estimates
Description:	Streamflow estimates from gauges (http://waterdata.usgs.gov/nwis/rt)
Restrictions:	None known

5	Downscaled Climate Model
Description:	Need downscaled climate model data for lake area (http://cida.usgs.gov/climate/gdp/)
Restrictions:	None known

Data Inputs – New Data Collection

1	Fish Habitat and Population
Description:	Determine population of trout by sampling the lake.
Data Management Budget:	20% of overall project budget, includes data storage, data organization support, integration, data visualization, backup, replication, and archive
Protocols:	Based on the California sport fishing protocol: http://oehha.ca.gov/fish/pdf/fishsampling121406.pdf
Quality Checks:	Comparison of results to previous studies. Spot checks against field manuals. Review of outliers.
Exclusive Use:	None
Restrictions:	None
Contact:	Jane Doe, jdoe@agency.gov

Data Outputs – Deliverables, Datasets and Products

1	Streamflow Estimates
Description:	Streamflow estimates for lake for the time period 2000-2010 and forward (using downscaled climate data)
Data Management Budget:	20% of overall project budget, includes data storage, data organization support, integration, data visualization, backup, replication, and archive
Format:	Csv
Quality Checks:	Scripts to review estimates and comparison to literature
Exclusive Use:	None
Restrictions:	None
Contact:	Jane Doe, jdoe@agency.gov

1	Trout Population Estimate
Description:	Estimates for trout population changes in the future based on streamflow, climate, and population changes.
Data Management Budget:	20% of overall project budget, includes data storage, data organization support, integration, data visualization, backup, replication, and archive
Format:	Shapefile (vector) and geotiff (raster)
Quality Checks:	Scripts to review estimates and comparison to literature
Exclusive Use:	None
Restrictions:	None
Contact:	Jane Doe, jdoe@agency.gov

A-3: Great Northern Landscape Conservation Cooperative Phenology Services

Data Inputs – Existing Data Collections

1	NASA Goddard Phenology
Description:	Sets of HDF files, 250m grids
Restrictions:	Distributed by NASA Goddard

Data Inputs – New Data Collection

No new collections will be developed.

Data Outputs – Deliverables, Datasets and Products

1	Phenology Geotiffs
Description:	Combined phenology geotiffs
Data Management Budget:	20% of overall project budget, includes data storage, data organization support, integration, data visualization, backup, replication, and archive
Format:	Geotiffs
Quality Checks:	Validation done via python scripts
Exclusive Use:	No exclusions
Restrictions:	No restrictions
Contact:	Robert Doe, rdoe@agency.gov

2	Dynamic Custom Visualization
Description:	Dynamic subsetting, charting, graphing, mapping, and reporting of phenology data
Data Management Budget:	Use 20% of overall project budget, includes data storage, data organization support, integration, data visualization, backup, replication, and archive
Format:	Java application (.java files)
Quality Checks:	Code review and regression testing
Exclusive Use:	No exclusions
Restrictions:	No restrictions
Contact:	Robert Doe, rdoe@agency.gov

Appendix B – Research Data Management Plan Example

Project Title: Wyoming Landscape Conservation Initiative Integrated Assessment

Data Inputs – Existing Collections

1	**Wyoming Geographic Information Center**
Description:*	This is a collection of over 100 natural resource geospatial datasets
Format:	Shapefile (vector) and geotiff (raster)
Source:	Various data contributors including: land management agencies (e.g., BLM), University of Wyoming, and local state agencies.
Data Processing & Scientific Workflows:	Data download and rasterizing
Backup & Storage:	Handled by Wyoming Geographic Information Science Center
Volume Estimate:	200 GB
Access & Sharing:	Data distributed by the Wyoming Geographic Information Science Center
Restrictions:*	Data are publically available with no restrictions.
Fees:	No fees
Citation:	Wyoming GISC Data Server http://wygl.wygisc.org/DataServer/

*Indicates that this information is also collected for the PDMP.

2	**Bureau of Land Management Disturbance Tracking System**
Description:*	Disturbance and reclamation areas associated with energy development on Bureau of Land Management areas
Format:	Shapefiles
Source:	Wyoming BLM State Office
Data Processing & Scientific Workflows:	Data download and rasterizing
Backup & Storage:	Handled by BLM
Volume Estimate:	10 GB
Access & Sharing:	Data distributed by BLM
Restrictions:*	No restrictions
Fees:	No fees
Citation:	Not known

Data Inputs – New Collections

No new data will be collected.

Software and Other Needs

1	**ArcGIS**
Description:*	Geographic Information Software for data processing
Restrictions:*	Proprietary/license software
Fees:*	USGS provides ArcGIS to USGS staff under an enterprise license agreement. All spatial processing can be effected using open source geospatial tools.
Source/Link:	http://www.esri.com

2	Python
Description:*	Scripting language
Restrictions:*	Open source software
Fees:*	No fees
Source/Link:	http://www.python.org

Data Outputs (e.g., Project Deliverables or Products)

1	WLCI Integrated Assessment
Description:*	The Integrated Assessment will result in a set of ESRI grids, shapefiles, and geodatabases that will allow land managers to assess the resource value associated with a cell or polygon. These assessments will cover the WLCI area, the southwest portion of the state of Wyoming.
Data Management Budget:*	The project will provide a total of $100K for data management for FY 11/12. This includes data storage, data organization support, integration, and custom presentation utilities along with data protection, storage, backup, replication, and archive.
Format:*	Final products will include geotiffs, shapefiles, and python scripts (.py files).
Data Processing & Scientific Workflows:	Sets of GIS data collected through collaborative efforts get partitioned into four categories (resource, condition, change agent, and future agent). The datasets are converted to 30m grids with a common origin, and then categorically combined. The values of each component grid get normalized, the combined grid converted to an index value, and the four categorical grids combined. Custom python scripts provide geoprocessing for resource assessment.
Quality Checks:*	The project provides a quality assurance team to validate all interim and final products. ScienceBase provides periodic checksum validation of all repository items, audits all inputs, and provides reviewable logs.
Backup & Storage:	The project uses ScienceBase for storage, replication, backup, and archive. This includes both interim and final products.
Metadata:	FGDC
Volume Estimate:	The final volume (source, interim, and final datasets, metadata, and processing scripts) will comprise less than 1TB.
Access & Sharing:	Prior to the completion of the project, only USGS staff has access to the data. The USGS requires review and approval of a publication or data series prior to wider distribution.
Exclusive Use:*	All project final and interim datasets will be available after completion of the data series publication. This will be complete at the conclusion of the project.
Archive Organizations:	The USGS Library will maintain the long-term repository for the publication, data series, and datasets.
Restrictions:*	All interim and final products will be released without restriction. Source data contains both proprietary and sensitive data.
Citation:	The USGS Library provides citation information.
Digital Object Identifier (DOI)/Link:	The USGS Library provides DOI/Links for project products.
Contact:*	Sarah Doe, sdoe@agency.gov

*Indicates that this information is also collected for the PDMP.