USGS National Wildlife Health Center

DRAFT Data Management Documents (for studies)

The NWHC Data and Information Management Planning Group started with the premise that NWHC staff are doing high quality work that has real and lasting value and as such it needs to be preserved, curated and promoted. That is really what a data and information management plan is about. It is a summary of how the data or information will be collected, preserved and shared during the lifecycle of that piece of work.

In learning about data management plans we found a University of Wisconsin presentation on the topic to be really helpful. We decided that if we could distill the information from that presentation down into a handful of questions it would capture the basic information needed for a plan at the study level. We came up with the idea of crafting an online template, pre-filling as much of the commonly used information as possible, so that a person could check a few boxes or select from a drop down menu to complete their plan. We pared our template down to 9 questions that cover the main components of a data and info management plan at the study level.

This, in conjunction with an over-arching 'master' data and information plan at the Center level which provides backup and additional detail will allow researchers/project leaders to craft study-specific plans quickly and easily, and comply with USGS/DOI guidelines. NOTE: USGS is also working on data management plans. We met with some folks working on that within our region. Bryan shared the draft USGS documents that came out recently and what we've drafted fits pretty well with what they have laid out. More work will be done to ensure the efforts meld appropriately.

Attached are several items:

Part 1 is the latest draft template of what will be an online fillable form for project leaders/lead investigators to fill out as they are developing their study plan.

Part 2 shows the template plus some draft responses

Part 3 shows a sample data management plan suitable for inclusion with a study plan.

Finally, the MasterDIMP Blueprint walks you through which template questions address specific components of a management plan. This is the outline for the Center-level plan identified above, which will complement regional and bureau plans.

Part 1: Simplified NWHC Data Management Template

Copy/paste Study Plan/BASIS Task Title here

(1) What data are you collecting or making?

- Observational
- Experimental
- Simulated

(2) What formats and standards are you using to collect and store your data?

- Documents: Word   PDF   text
- Databases: MS Access   SQLServer   R   text   other
- Other: Excel

(3) Are you using any non-NWHC standard terminologies (click for list)

- Yes   (please enter data category here) _____
- No

(4) Do you have

a) a data dictionary? *** define term **(If not, help is available from ...**

b) Metadata? *** define, + help available from ...

c) Methods documentation for any specialized data manipulations ONLY (such as ?

d) does not involve data analysis (information product)

(5) Are there any specific requirements for your project/data (maybe because of interagency data gathering, etc.)?

- Yes _____
- No

   Suggestion: consider funder requirements, any requirements of other agencies involved, confidentiality, security.

(6) Where will you store your data short-term?

a) NWHC network drive, with backups on local hard drive and external hard drive/flash drive
b) NWHC local hard drive, with backups on external hard drive or flash drive
c) Other - please start with the words "Data will be stored …"_____

Suggestion: best practices suggest you keep 2-3 running copies of your data during the life of the project. If one of those copies is on a network drive, you ensure that the data will be backed up nightly and stored offsite weekly in case of disaster.
Suggestion: think about how fast your data will grow over time and how frequently it will change. Do you need backups daily, or more/less frequently? For more frequent backups, see the IT group for assistance as needed.

(7) Where will you store your data long-term?

a) Data repository !!
b) NWHC network drive
c) Other (please explain)_____

Suggestion: best practices suggest you consider the lifetime of your data. NWHC may develop its own data repository; if not, the MWR/USGS will definitely have various repositories around. Using a repository ensures your data are preserved, migrated as needed to up-to-date, accessible formats, etc. It also ensures your data are secure and backed up religiously so you do not have to worry. Putting your data in a repository does not imply sharing rights, just preservation.

(8) Once you're finished with the project, what do you think people should be able to do with your data?

a) View metadata online, contact PI for more information
b) View raw/derived data online, contact PI for more information
c) View/download metadata, full documentation and data following posted data sharing plan
d) Other _____

NOTE: This will follow guidelines in the NWHC data sharing policy; you retain first rights to your data for a specified period of time. What happens afterwards? You cannot copyright facts or any USGS data. What would someone unfamiliar with your data need in order to find, evaluate, understand, and reuse them? Two parts: metadata and methods

(9) Are there any specific issues regarding your data not covered in the NWHC Data and Information Managemnt Plan?

- **Yes** _____
- No

Part 2: Template plus boilerplate responses (draft)

Copy/paste Study Plan/BASIS Task Title here

(RESP) Experimental Analysis of the ... (or whatever the title is)

(1) What data are you collecting or making?

- Observational
- Experimental
- Simulated

(RESP) This study will collect _____ data following the methods described in the attached proposal/study plan.

(2) What formats and standards are you using to collect and store your data?

• Documents:	Word	PDF		text	
• Databases:	MS Access	SQLServer	R	text	other
• Other:	Excel				

(RESP) Data will be stored primarily in _____ format, and

(3) Are you using any non-NWHC standard terminologies (click for list)

- Yes   (please enter data category here) _____
- No

(RESP)
(if yes). will use some standard terminologies documented in the NWHC master data and information blueprint, as well as others unique to this study. Specialized terminology documentation will be included in the data dictionary files noted below; information on standard terminologies is available from the NWHC master data and information blueprint.

(if no). will use standard NWHC terminologies documented in the NWHC master data and information blueprint.

(4) Do you have

a) a data dictionary?   *** define term **(If not, help is available from ...**

b) Metadata?   *** define, + help available from ...

c) Methods documentation for any specialized data manipulations ONLY (such as ?

d) Does not involve data analysis (information product)

(RESP)
(if yes to a). All data fields have been identified and documented in a data dictionary,

(if no to a). A study-specific data dictionary is under development,

(if yes to b). FGDC standard metadata have been developed ,

(if no to b). FGDC standard metadata will be developed/amended over the course of the study,

(if yes to c). and the following data techniques will be used . [insert user text here].

(if no to c). [insert 'and' between answers to 4a and 4b and put period at end of sentence.]

(if yes to d). ...

(5) Are there any specific requirements for your project/data (maybe because of interagency data gathering, etc.)?

- Yes   (please enter info here in sentence form) _____
- No

    Suggestion: consider funder requirements, any requirements of other agencies involved, confidentiality, security.

(RESP)
(if yes). [info direct from user input]
(if no). [nothing]

(6) Where will you store your data short-term?

a) NWHC network drive, with backups on local hard drive and external hard drive/flash drive
b) NWHC local hard drive, with backups on external hard drive or flash drive
c) Other - please start with the words "Data will be stored …"_____

    Suggestion: best practices suggest you keep 2-3 running copies of your data during the life of the project. If one of those copies is on a network drive, you ensure that the data will be backed up nightly and stored offsite weekly in case of disaster.

Suggestion: think about how fast your data will grow over time and how frequently it will change. Do you need backups daily, or more/less frequently? For more frequent backups, see the IT group for assistance as needed.

(RESP)

(If a). Data will be stored following USGS/NWHC best-practices guidelines. The master copy will reside on the NWHC local area network, with full backup support and redundant offsite storage provided by local IT staff. Data will be updated on the master network copy; file(s) will be copied to the principal investigator's local hard drive and to an external hard drive/flash drive for additional security at least weekly by project personnel.

(If b). Data will be stored on the principal investigator's local hard drive and to one or more external hard drives/flash drives for additional security. The master copy is considered the copy on the principal investigator's local hard drive, and project personnel are responsible for regular backups of the data, and copies to external drives for additional security at least weekly. The principal investigator will determine the backup/copying frequency.

(If c). [copy info directly from user input]

(7) Where will you store your data long-term?

    d)   Data repository !!
    e)   NWHC network drive
    f)   Other (please explain)_____

> **Suggestion: best practices suggest you consider the lifetime of your data. NWHC may develop its own data repository; if not, the MWR/USGS will definitely have various repositories around. Using a repository ensures your data are preserved, migrated as needed to up-to-date, accessible formats, etc. It also ensures your data are secure and backed up religiously so you do not have to worry. Putting your data in a repository does not imply sharing rights, just preservation.**

(RESP)

(If a). Data will be preserved long-term following USGS/NWHC best-practices guidelines. Once the study is complete and data have been analyzed/published/_____, the principal investigator will authorize transmission of the data sets and documentation to a USGS/MWR data repository.

(If b). Data will be preserved on a NWHC network drive until such time as the principal investigator authorizes ....

(If c). [copy info directly from user input]

(8) Once you're finished with the project, what do you think people should be able to do with your data?

      e)   View metadata online, contact PI for more information
      f)   View raw/derived data online, contact PI for more information
      g)   View/download metadata, full documentation and data following posted data sharing plan
      h)   Other _____

> **NOTE: This will follow guidelines in the NWHC data sharing policy; you retain first rights to your data for a specified period of time. What happens afterwards? You cannot copyright facts or any USGS data. What would someone unfamiliar with your data need in order to find, evaluate, understand, and reuse them? Two parts: metadata and methods**

(RESP)

(If a). Metadata for this study will be stored on the NWHC website and on various clearinghouse sites. Users interested in obtaining raw/derived data or additional information on the study may contact the principal investigator.

(If b). Metadata for this study will be stored on the NWHC website and on various clearinghouse sites; data will be available for download from the NWHC website following _____.
Users interested in using the data or in additional information on the study should contact the principal investigator.

(If c). Metadata for this study will be stored on the NWHC website and on various clearinghouse sites; data and complete documentation will be available for download from the NWHC website following _____ and may be used with appropriate citation.

(If d). [copy info directly from user input]

(9) Are there any specific issues regarding your data not covered in the NWHC Data and Information Managemnt Plan?

      • **Yes** _____
      • No

(if yes). [copy info directly from user input]

Part 3: Sample Study Data Plan

Title: Evaluation of the effectiveness of various mouse designs in retarding development of carpal tunnel syndrome in heavy-duty computer users.

This study will collect observational data following the methods described in the attached proposal/study plan.   Data will be stored primarily in MS Access format, and will use standard NWHC terminologies documented in the NWHC master data and information blueprint.   A study-specific data dictionary is under development, and FGDC standard metadata will be developed/amended over the course of the study.

Data will be stored following USGS/NWHC best-practices guidelines. The master copy will reside on the NWHC local area network, with full backup support and redundant offsite storage provided by local IT staff. Data will be updated on the master network copy; file(s) will be copied to the principal investigator's local hard drive and to an external hard drive/flash drive for additional security at least weekly by project personnel.

Data will be preserved long-term following USGS/NWHC best-practices guidelines. Once the study is complete and data have been analyzed/published/_____, the principal investigator will authorize transmission of the data sets and documentation to a USGS/MWR data repository.

Metadata for this study will be stored on the NWHC website and on various clearinghouse sites; data and complete documentation will be available for download from the NWHC website following _____ and may be used with appropriate citation.

Detailed documentation on standard practices, policies, and guidelines for data and information management at USGS/NWHC may be found in the NWHC master data and information blueprint. Please contact _____ for a copy.

NWHC 'Master' Data and Information Management Blueprint

Introduction

Effective data and Information management will increase the availability, visibility and use of NWHC data. Properly managed, it will save time, simplify the lives of those generating data by providing tools and assistance for managing data, and ensure the preservation and protection of NWHC data and information over time.

Purpose

This plan describes the data and information collected by NWHC personnel in general terms, and identifies the components of a Data and Information Management Plan (DIMP), which includes how the data and information will be processed, managed, preserved and shared throughout the life cycle. It provides standard information on what categories of data and information will be collected, and specifically covers the topics of audience, sources, standards, policies, authorities and mandates applicable to all NWHC work. This over-arching framework provides backup for all types of NWHC data management plans, including those for study plans and any data and information products Center staff produce, such as fact sheets or wildlife health bulletins. Individual data management plans for studies/products will reference this document in lieu of duplicating information that is applicable across NWHC to all studies and products. In addition to this document, NWHC is providing principal investigators (PIs) with a 'fillable' template consisting of 9 questions with standard 'boilerplate' text options for use in individual data management plans PIs may select one of the standard options for each question or develop their own text.

Data and Information Life Cycle (components of a data and information management plan)

I.   Plan (Title and Questions 5 & 9 in the template/fillable form)

This section includes basic information on the purpose of the data collection, the audience for the data, and the sources of the data. It also includes information on standards, policies and authorities relevant to data management in NWHC/USGS/DOI.

a) WHY. This master blueprint will have some general information here on why NWHC collects data, and our audience.

b) WHAT AUTHORITY. This master blueprint will have boilerplate information in this section on standards/policies/ authorities that can be referenced for individual plans. This information will not have to be in each individual plan but there will be a line inserted at the end of the individual plan saying 'see master plan for more information on data management at NWHC' or something similar.

c) HOW USED. This section of the master blueprint will include how the data will be used and who has what responsibilities for the data, and any interagency agreements that address data property rights.

d) On the fillable form, PIs can copy/paste the title and/or summary narrative from their BASIS task or study plan to the first open area. Additionally, any special or non-standard issues will have to be added by the PI (reference Question 9 on fillable form)

Standards are linked to the 'describe' and 'manage quality' elements below.

II.   Acquire (Questions 1 & 2)

This section includes: How will the data be collected? If any data are coming from another entity, is there detailed metadata available?   Will data gatherers/processors need any training on how to collect the data properly and transmit it to a central point? Or for anything else?

a) This master blueprint will have information here on the 3 main types of data (observational, experimental, and simulated) with standard verbiage for each. (Will be pretty general)

b) PIs can check an option on the template for type of data (Question 1), and select formats (Question 2). There will be an option for PIs to enter any additional information desired.

III.   Process (Questions 3, 4, 6)

This section builds upon the last section but includes any changes made to the original data as part of the study. It also includes the temporary storage of the data while they are collected/analyzed/etc.

a) This master blueprint will have information on NWHC 'standard' data dictionaries, metadata standards and methods documentation which can be referenced instead of re-copied (Question 4).   PIs will have to document use of non-standard terminologies (Question 3) in their data dictionary/methods.

b) Question to think about (to be posed somewhere) – what would someone unfamiliar with your data need in order to find, evaluate, understand, and use them?

c) PIs can select from an NWHC standard option for short-term data storage with one click (Question 6) which will insert 3 short paragraphs covering data backups and data security into their plan, or they can write their own paragraph(s) describing how the data will be backed up and secured.

d) PIs can reference SOPs or study plans as well

e) Version control for info products?

IV.   Analyze (Question 4, revisited)

This section includes methods for data analysis and is specific to each dataset. The master plan will have some general information saying that data analysis follows DOI/USGS policies ... but

expect to have each study reference an existing study plan or other study document for methods.

a) For studies that do specialized data manipulations this will have been covered in Q4c. For information products, there will be an option/check box for 'does not involve data analysis'.

V.   Preserve (Question 6, revisited)

This section includes issues of data security (confidentiality, integrity, availability) as well as backups and duplicates copies (linked to 'backup and secure' element).

a) This master blueprint will have boilerplate information on data repositories and keeping data accessible (succession planning), the USGS records schedule-> disposal and/or archiving here.

b) PIs have already selected an option that will cover the basics of this area.

c) ? Other security issues?

VI.   Publish/Share (Questions 7 & 8)

The section addresses how and when NWHC will share data. This master blueprint will reference the NWHC data and sample sharing policy and re-iterate embargo periods, etc.

a) Boilerplate options for internal access/partner access/public access will be available in this master plan and will include the user interfaces for data access and how users may request assistance/support.

b) PIs can select from several options on where data will be stored long-term (Q7) or write their own statement.

c) PIs can select from several options (Q 8) in terms of metadata and/or data shared or write their own statement

Includes standard data accessibility and data formats (linked to all 3 cross-cutting elements)

Cross-Cutting Elements: These are questions/issues for PIs and other data gatherers/consolidators to consider before writing or filling in a data management plan. They are intended to stimulate the user to look at the importance and usefulness of the data long-term.

I.   Describe

Questions: What types of data are being collected? How much will you have and how fast will the dataset grow? How often will the data change? Impacts #/timing of backups, etc.

Includes data definitions and naming conventions, data formats (paper/electronic/ platform/etc.), metadata standards, requirements for data dictionaries and methods description.   Do not forget to include UNITS.

II.   **Manage Quality**

Includes quality assurance and QA/QC procedures. This master blueprint will have standard terminology on procedures across NWHC.

Includes standards – in terms of 'if it's there already don't re-invent it'; sharing relies on standards! "Keep it Simple".

III.   **Backup and Secure**

First question: how valuable is your data? How much would it cost to recreate? Is it possible to re-create it?

Second question: how long do you need to keep the data (ties back to USGS records management requirements).

****

Draft for master plan:

Best practices suggest keeping at least 3 copies of all datasets, both while under development and when final. NWHC best practices suggest keeping the following: one copy on the NWHC network, one copy on the principal investigator's local drive, and one copy on a USGS flash drive or external hard drive.

All changes should be made to one copy of the data, and then the entire dataset should be replicated from this 'primary' or master copy to the other copies. In general, the network copy – since it is accessible to all members of the study team – is considered the primary copy. It is the responsibility of the principal investigator to determine how and how often the copies need to be synchronized -- based on the expected growth rate of the data and the amount of work needed to re-create the data should it be lost -- and to determine if the standard NWHC backup procedures documented below are sufficient to ensure the security of the data.

NWHC network backup procedures. For complete details, please see SOP (IT-SOP-016) in Appendix A. Backup tests are performed quarterly. Network backups performed by NWHC IT staff include the following:

1) hourly copies of all user data on shared drives to a local storage array
2) nightly tape backups (full backup, not incremental) each workday at 9pm of all user data on shared drives.
3) daily backups of all user data on shared drives to a remote hard drive in Ann Arbor, MI (intermittently right now but working on it).

Data stored on the network and on campus workstations may normally be stored unencrypted; USGS policies discourage storage of sensitive information on shared drives. Encryption options for mobile devices are available following standard USGS policies; contact the IT group for purchasing information. NWHC does not recommend or encourage compression of any dataset.

Appendix A

IT-SOP-016

....

Terminology and Other Standards

Metadata –

NWHC metadata will follow the FGDC standard in compliance with Executive Order 12906: COORDINATING GEOGRAPHIC DATA ACQUISITION AND ACCESS: THE NATIONAL SPATIAL DATA INFRASTRUCTURE, signed by President Bill Clinton on April 11, 1994; as amended by President George W. Bush by issuance of Executive Order 13286 on March 5, 2003.

Assistance with metadata development for NWHC projects is available through Core Sciences.

Location Names –

Note to NWHC users – if you use the standard Center lookup tables for lCounties, lState and lNation, (on the CONIFER drive) you may assume that you are complying with federal government standards and that the information will be reviewed at least annually.

NWHC does not currently maintain a lookup table for other named geo-locations, but allows free text entry of specific location information as needed. Users are strongly encouraged to use the USGS website (geonames.usgs.gov) identified below to follow accepted usage guidelines.

> INCITS 38 – 2009. Codes for the Identification of the States and Equivalent Areas within the United States, Puerto Rico, and the Insular Areas. Replaces FIPS 5-2, withdrawn September 2, 2008.

> INCITS 31 – 2009. Codes for the Identification of Counties and Equivalent Areas of the United States, Puerto Rico, and the Insular Areas. Replaces FIPS 6-4, withdrawn September 2, 2008.

> INCITS 454 – 2009. Codes for the Identification of Metropolitan and Micropolitan Statistical Areas and Related Statistical Areas of the United States and Puerto Rico. Replaces FIPS 8-6, withdrawn February 8, 2005.

> INCITS 455 – 2009. Codes for the Identification of Congressional Districts and Equivalent Areas of the United States, Puerto Rico, and the Insular Areas. Replaces FIPS 9-1, withdrawn February 8, 2005.

> ---   http://geonames.usgs.gov/domestic/index.html. Named populated places, primary county divisions, and other locational entities of the United States, Puerto Rico, and the Outlying Areas. Replaces FIPS 55- DC3, withdrawn September 2, 2008.   Searchable.

ISO 3166 – Country Codes.   http://www.iso.org/iso/country_codes/country_codes. A US profile is under development, but currently USGS is following the ISO standard except where it conflicts with INCITS codes. Replaces FIPS 10-4, withdrawn September 2, 2008.

Georeferencing –

Projection and Coordinate System
All digital geospatial data should reference the coordinate system appropriate for it's use and it should be documented in the metadata. All spatial data collected shall be geo-referenced and provided in a standard projection. Digital geospatial data should be referenced to two coordinate systems--the current standard system used by the individual park (generally UTM, NAD83) and a regional-scale system (Geographic, NAD83). The steps used to get the data into the proper projection must be documented in the metadata. The project manager must specify, approve and document any deviation from these projection standards.

Data Standard
The standard projection is geographic with the following parameters as per Executive Order 12906 (http://www.fgdc.gov/publications/documents/geninfo/execord.html) and the Federal Geographic Data Committee (FGDC) standards:

Datum North American Datum 1983
Spheroid GRS 1980
Units Decimal Degrees

NEED USGS Specifics

Horizontal / Vertical Accuracy and Precision
All spatial data collected shall be analyzed for their spatial accuracy and shall meet or exceed the National Map Accuracy Standards for the particular scale intended (for more information see http://rockyweb.cr.usgs.gov/nmpstds/nmas.html). Longitude and Latitude coordinates for geographic data should be recorded to a minimum 5 significant digits to the right of the decimal point and stored in double precision attribute or database fields. Any calculations done with location data should be done at double precision with the results rounded or truncated to the appropriate propagated error limits. All calculations and processing completed on the spatial data shall be reported in the metadata.

Species –

Note to NWHC users - if you use the standard Center lookup table, lSpecies, you may assume that you are complying with federal government standards and that the information will be reviewed at least annually. The lSpecies table links common name and standard taxonomic information (from class to species) from IT IS to species status, TSN number, AOU code (avian only). Species status are checked using CITES.

--   http://www.itis.gov   Integrated Taxonomic Information System. USGS is a founding member and ongoing collaborator in the maintenance of this system.

31 Jan 2014