4-2015

Data Management Plan Writing Ready Reference

Kiyomi D. Deards
University of Nebraska-Lincoln, kdeards2@unl.edu

DeeAnn Allison
University of Nebraska-Lincoln, dallison1@unl.edu

Data Management Plan Writing Ready Reference

ACRL e-Learning

Kiyomi D. Deards and Dee Ann Allison University of Nebraska – Lincoln

kdeards2@unl.edu, dallison@unl.edu

Updated April 6, 2015

Table of Contents

Writing Data Management Plans Checklist

1. Don't panic, breath!

2. Schedule a meeting to discuss the project and data management needs, allow at least an hour and a half. Average length (15 minutes- 2 hours).

3. If possible, everyone should have a copy of the grant proposal/project outline. Examples of experts to include in the discussion: librarians for related disciplines, metadata specialists, and digital imaging and archives specialists.

3.5 Ask if there are going to be any restrictions to access the data (copyright, patent, trademark, privacy, national security, Export Control, etc.).

4. Send the researcher some basic information on data management plans to review before your meeting. This can be a handout you've created, a web page on data management, and copies or links to sample data management plans.

5. Print out 3 different examples of data management plans that provide various approaches to data management so that you can refer to them during your consultations. (Single PI, Group, International, Interdisciplinary, small data sets, large data sets, physical sample, strictly computational etc.).

6. Go to the consultation. Explain metadata, image standards, open file formats, trusted repositories, standard data backup requirements (3 copies minimum, at least one in a

geographically separate location). Emphasize the need to address every file type and file conversion, as well as physical samples, physical laboratory notebooks, and other methods of data collection such as images, videos, etc. Funders can have multiple levels of direction which may be changed at any point in time. Ex: NSF has overall guidelines, plus there are directorate specific and program specific guidelines. Grantees will have to review all of these to ensure their data management complies every time they submit a proposal.

6.5 If the researcher would like you to manage their data they should write your time into the grant. Unless you are already working for the group they will most likely have to buy out your time since managing data for a major research project is a usually a full time project.

7. Review existing data repositories to recommend an appropriate repository (if one exists). Check the index of the Registry of Research Data +Repositories (1. by subject, 2. by content type and 3. by country): http://www.re3data.org/browse/. If a researcher does not have an existing subject repository, and there is no institutional repository available, FigShare may be a viable alternative. http://figshare.com/

8. If your institution or consortia manages a local repository provide information on policies, requirements, costs and procedures for depositing data.

9. Review the plan. If something confuses you it will confuse a reviewer. Ensure that all samples, data file types, and methods of recording data have been thoroughly addressed. Send your detailed comments to the researcher; if possible have another librarian review the plan and add their own comments.

10. After a successful consultation ask the researcher if their plan can be shared as a training tool within your organization.

11. Relax, you've done it!

Table 1. Open Formats That Are Not Proprietary

Open Format	Description
7z	archiving/compression
ABW	text - Abiword
Aften	Lossy audio codex
ALAC	lossless audio codec
APNG	animated PNG
Apple Lossless	Lossless audio codec
ASCII	text
BonkEnc	hybrid lossy/lossless codec
bzip2	compression
CELT	Lossy audio compression
CMML	timed metadata
CSS	stylesheet
CSV	comma separated fields
DAISY	talking book format
dcaenc	Lossy audio codex
DjVu	scanned images
DVI	text - device independent
EAS3	floating point data
ELF	executable and linkable format
ePub	open e-book
Ffmpeg	video codec
FictionBook	open e-book - XML
FLAC	lossless audio codec
FreeOTFE	encrypted data container
GIF	Lossless image Format
GPX	GPS, waypoints, tracts & routes
gzip	compression
Hierarchical Data Format	multidimential array structures
HTML	text for Web pages
Huffyuv	lossless video codec
iCalendar	calendar data format
IFC	data modeling for construction
JPG	lossy image format
JPEG 2000	lossy or lossless image format
JSON	object notation
Kvazaar	video encoder
Lagarith	video codec
LAME	Lossy audio compression

LaTeX	text markup language
libgsm	Lossy audio codex
libvorbis	Lossy audio compression
libvpx	video codec
LTFS	Linear Tape File System
lzip	compression
MAFF	archiving
Matroska (mkv)	container for multimedia formats
MNG	moving pictures
MP3	audio format
Musepack	Lossy audio compression
NetCDF	scientific data
NZB	multipart binary files
Ogg	container for lossy audio formats & video formats
SXI	text - Open Office, Neo Office and Star Office
OpenAVS	audio codec
opencore-amr	Lossy audio codex
ODP	open document format- Open Office, Neo Office, Star Office
OpenEXR	image file format
OpenH264	video codec
OpenXPS	text
Opus	Lossy audio compression
PAQ	compression
PDF subsets	some PDF subsets are open
PNG	lossless image format
PostScript	text
RDF	Resource Description Framework
RSS	news syndication
RFT	unformated text
SDXF	Structured Data eXchange Format
SFV	checksum format
SMIL	media playlisting format and integration language
Speex	speech codec
SQX	archiving/compression
STW,SXW	text -Star Office, Open Office and Neo Office word
SVG	vector image format
tar	archgiving
TeX	formated text suppots math & science
Theora	lossy video compression format
TooLAME	Lossy audio compression
TrueCrypt	container for encrypted data

TTA	lossless audio compression
UTF-16	16 bit text
UTF-8	8 it text
Vorbis	lossy audio compression format
VRML/X3D	3D data formats
WavPack	hybrid lossy/losslessaudio codec
WebDAV	Internet filesystem format
WebM	video/audio format
WebP	image format
WMV	windows media file
XHTML	text - xtensible HyperText Markup Language
XMF	music file format
XML	text markup language
XSPF	playlist format for multimedia
Xvid	video codec
xz	compression
YAML	human readable data format
ZIP	archiving/compression

Table 2. Closed Data Formats

Closed Format	Description
CDR	CorelDraw's native format
DWG	AutoCAD
PSD	Adobe Photoshop lossy or lossless
WMA	Microsoft audio format
LIT	Microsoft ebook reader
mus	Music format
PDF	Adobe text
iwork	Apple text
doc, docxdot,dotx,,pps,xls	Microsoft text, presentation and spreadsheet
AIFF	Apple uncompressed audio
WAV	Microsoft lossless audio format
JPEG	Lossy - sometimes considered open format
tiff	lossless compressed image format
PAGES	Apple text document
WPD	Word Perfect
WPS	Microsoft Works
WRI	Microsoft write
MOV	Apple Quicktime video
SWF, FLA	Adobe Flash

Major Federal Funders and Other Granting Agencies

Aggregators:

Grants.gov http://grants.gov/

The Foundation Center http://foundationcenter.org
(Each state should have at least one location with free full access, every can access partial information.)

Fundsnet Online Services http://fundsnetservices.com

Google http://www.google.com

Select Major Federal Funding Agencies

CDC – Center for Disease Control/Prevention
http://www.cdc.gov/about/business/funding.htm

DoE – Department of Energy

http://www.energy.gov/public-services/funding-financing

Department of Education http://www.ed.gov/fund/landing.jhtml

DoD – Department of Defense http://grants.gov/

NEA – National Endowment for the Arts http://arts.gov/grants

NEH – National Endowment for the Humanities http://www.neh.gov/grants

NIH – National Institutes of Health http://grants.nih.gov/grants/oer.htm

NSF – National Science Foundation http://www.nsf.gov/funding/

USDA NIFA – National Institute of Food & Agriculture

http://www.csrees.usda.gov/business/business.html

Data Repositories

Registry of Research Data +Repositories http://www.re3data.org/browse/
This list is partially curated, reviewed resources have a green check symbol.
Search by: subject, content type, or country

FigShare http://figshare.com – Free and open to all researchers, ideal for researchers with small to medium amounts of data and those with no domain or institutional repository.

DataBib – http://databib.org/ a searchable bibliography of research data repositories. (Please note this is an unevaluated list and is merging with R3Data http://www.re3data.org/.)

Two most mentioned repositories:

Dryad – http://datadryad.org/ A specialized repository of scientific and medical data, primarily data associated with peer-reviewed journals. See http://datadryad.org/pages/integratedJournals for deposit costs for non-affiliated researchers and sponsored journals.

ICPSR (Interuniversity Consortium for Political and Social Research) – http://www.icpsr.umich.edu/icpsrweb/deposit/index.jsp A large social and behavioral sciences data repository. You will note that open responses are generally not available for downloads.

Good and Bad Examples from DMPs

Good Examples from DMPs

Example 1:
This Data Management Plan (DMP) covers the data which will be collected for a longitudinal study at { }. The study projected to be conducted between {} and {}. The study will collect non-sensitive data from a cohort of mixed-gender subjects ranging from 18 to 85 years of age. No other personal identifiers will be collected during the study apart from those identified above. Subjects who consent to participate will be anonymously surveyed by asking to respond to a simple yes/no, single question verbal survey. The results will be manually recorded by the surveyor. The data collected during this study will be archived with { }. The data will be stored in a specific virtual archive and will be made publicly available through {}. This {} data archive is a well-established and trusted archive in the social science field. As a member of the Data Preservation Alliance for the Social Sciences (Data-PASS) and the Library of Congress National Digital Stewardship Alliance (NDSA) {} provides a strong archival and data distribution resource to the project. The aim and purpose of this DMP is to detail and guarantee the preservation of the data collected during this study, as well as any results derived from the associated research. This DMP is intended for review by relevant NSF personnel, as well as {} staff and {} affiliated directly with this study and the collection and preservation of the associated data and research.
Comprehensive institutional and research group guidelines specified by {} were applied regarding the collection of this data.

This study will only collect non-sensitive data. No personal identifiers will be recorded or retained by the researchers in any form. There are no copyright or licensing issues associated with the data being submitted. The data being submitted will be made publicly available through the {} by {}. There will be no additional restrictions or permissions required for accessing the data. Findings will be published by the researchers based on this data; the estimated date of publication is { }. There is an agreement regarding the right of the original data collector, creator or principal investigator for first use of the data. The specified embargo period associated with the data being submitted extends from the projected conclusion date for initial research until six months after projected publication date for the findings. The embargo will be lifted by {}.
The associated data types will be captured using Qualtrics survey software and analyzed using SPSS data analytics tools. The researchers are not aware of any issues regarding the effects or limitations of these formats regarding the data being submitted.

General metadata related to the survey topic will be created for the data being submitted. The associated metadata will be manually created in XML file format. DDI metadata standards will be applied during the creation of the metadata.

At {} experience with, and commitment to, secure data archiving is well established and is in keeping with established {} Information Security Policies. During the implementation of the survey, associated research data will be physically stored on a password-protected secure server maintained by {} using standard SPSS file formats. No data will reside on portable or laptop devices, and no other external media/format(s) will be used for data storage. Research data is backed up on a daily basis. The researchers are currently responsible for storage, maintenance

and back-up of the data. The specific storage volume of the data being submitted will be not more than 1GB maximum. The long-term strategy for the maintenance, curation and archiving of the data will be implemented when the data and associated research are migrated to {} for archiving using the {}. Preservation, review and long-Term management of data collected during this study will be archived with []. The data will be stored in a specific virtual archive and will be made publicly available through the []. As a result of this arrangement, there are no specific financial considerations of which the researchers are currently aware which might impact the long-term management of the data. The research and archival staff of the {} will review this DMP upon accession of the data in order to ensure and demonstrate compliance. The DMP will again be reviewed by {} and archival staff prior to ingest and release into the {}.

What is good about this example?

1. Notice how specific the researchers are about roles and responsibilities.
2. Privacy and licensing, etc. is addressed.
3. Procedures for handling data are discussed.
4. Specific are given for long-term archiving of the data and when it will become public.
5. Information is provided on the credentials for the archive site.

Example 2.

The project will leverage existing metadata standards currently stored in Ecological Metadata Language (EML) format. We will add additional metadata entries for the arthropod community composition and arthropod stoichiometry; field notes taken during the time of collection will be recorded. Morpho software will be used to generate the metadata file in EML. We chose EML format for our metadata since it allows integration with existing NutNet data housed in the Knowledge Network for Biocomplexity (KNB) data repository. After publication of manuscripts based on the data we collect, we will share our data and metadata with the NutNet community via data updates sent annually as csv files from the existing central relational database. We will also submit both of our datasets (abundance and stoichiometry) to the {}, an archive for digital preservation. This will occur within a year of publication. The data will be publicly available via the Digital Conservancy, which provides a permanent URL for digital documents. 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 participating institutional and NSF policies. Depending on such policies, materials may be transferred to others under the terms of a material transfer agreement. Those that use the data (as opposed to any resulting manuscripts) should cite it as follows: {} [URL]; accessed on ddmmyyyy. This information will be described in the metadata.
Intended and foreseeable users of the data are NutNet collaborators and participants, as well as other scientists interested in arthropod plant relationships. This data set could be used in combination with similar data sets from other NutNet sites or for meta-analysis. We will preserve both arthropod datasets generated during this project (abundance and stoichiometry) for the long term in the {}. We will include the .csv files, along with the associated metadata files. We will also submit an abstract with the datasets that describe their

original context and potentially relevant project information. {} will be responsible for preparing data for long--term preservation and for updating contact information for investigators.

What is right from this example?

1. Details metadata standard and interoperability of metadata.
2. Specifies what and when data will be shared and format of data. Format in this case is important since this is being stored in a database.
3. Specified when data will be made publically available.
4. Provides for permanent URL for data.
5. Stipulates additional possible distribution of data.
6. Specifies data citation.
7. Describes possible circumstances for data re-use.
8. Defines roles and responsibilities.

Example 3:

I will share phenotypic data associated with the collected samples by depositing these data at {}, which is an NIH-funded repository. Genotype data will be shared by depositing these data at {}. Additional **data documentation and de-identified data will be deposited** for sharing along with phenotypic data, which includes demographics, family history of XXXXXX disease, and diagnosis, consistent with applicable laws and regulations. I will comply with the NIH GWAS Policy and the funding IC's existing policies on sharing data on XXXXXX disease genetics to include secondary analysis of data resulting from a genome wide association study through the repository. Meta-analysis data and associated phenotypic data, along with data content, format, and organization, will be available at {}. Submitted data will confirm with relevant data and terminology standards. I agree that data will be deposited and made available through {}, which is an NIH-funded repository, and that these data will be shared with investigators working under an institution with a Federal Wide Assurance (FWA) and could be used for secondary study purposes such as finding genes that contribute to process of XXXXXX. I agree that the names and Institutions of persons either given or denied access to the data, and the bases for such decisions, will be summarized in the annual progress report. Meta-analysis data and associated phenotypic data, along with data content, format, and organization, will be made available to investigators through {}. I agree to deposit and maintain the phenotypic data, and secondary analysis of data (if any) at {}, which is an NIH-funded repository and that the repository has data access policies and procedures consistent with NIH data sharing policies. I agree to deposit genetic outcome data into {} repository as soon as possible but no later than within one year of the completion of the funded project period for the parent award or upon acceptance of the data for publication, or public disclosure of a submitted patent application, whichever is earlier. I agree that I will identify where the data will be available and how to access the data in any publications and presentations that I author or co-author about these data, as well as acknowledge the repository and funding source in any publications and presentations. As I will be using {}, which is an NIH-funded repository, this repository has policies and procedures in place that will provide data access to qualified researchers, fully consistent with NIH data sharing policies and applicable laws and regulations.

What is right with this example?

1. Data is stored in a federal repository.
2. Specifies regulatory compliance.
3. Specified who has access and possible re-use of data.
4. Specified when data will be made available to others.
5. De-identifies data to insure privacy. (Doesn't say is in compliance with HIPAA)
6. Specified additional documents will be provided (but doesn't specify much about the metadata, which would have made this stronger)
7. Although this was provided as an example by NIH in 2010, it lacks some of the details that we would recommend, including more information about formats, and metadata. However, because the data is going into a NIH repository, that kind of information may be not be necessary in a grant to NIH. This is a case where I would defer to the researcher's expertise and knowledge of NIH practices and a reminder to review the latest requirements for data management plans on a funder's website.

Bad Examples from DMPs

Example 1.

All sample data will be collected and organized using [Speciality Software Name]. The files will contain information about sample characteristics and the conditions under which these characteristics were measured. Approximately 1-2 Gb of data will be generated.

What is wrong with this example?

1. No files formats are mentioned, nor are preservation issues of the file formats.

2. Sample characteristics are not defined. There may not be room in the DMP to do so, but it should at least refer back to the section of the grant which defines them.

3. Is the data being generated text, numeric, graphs, etc.? If data is out in a common format, Word, Excel, TXT, CSV, JPG, TIFF, XHTML, MPG, etc. you do not need to state this. Many scientists have specialty code written to run their custom built instruments and may not generate data in standard formats (usually these can be converted into CSV and/or TXT).

Example 2.

All files will be stores on the PI's secure computer. All laboratory notebooks will be stored in the PI's office.

What is wrong with this example?

1. There is no mention of backup copies. A minimum of three copies with at least one copy in a geographically separate location are required to consider data securely backed up.

2. It should also mention the procedure for data backup durring and after the project which should include not only the physical steps taken but also a timeline. Ex. Every week the PI will back up the data on an external hard drive stored at the PI's home. Every night data on the PI's

computer is automatically backuped at midnight by the university's Information Services Department.

3. How are the laboratory notebooks stored? Are they in a box on the floor (risk of flooding), on a shelf, or in a cabinet (locked or unlocked) when not in use? Who can access the PI's office and would the notebooks be secure when the PI is not there?

Example 3.

Data will be available to anyone who desires access to our data. When possible, data will be made available online.

What is wrong with this example?

1. Does not state how “anyone” will be able to access the data.

2. Does not state how online access will be provided.

3. Does not state that because the data contains no restricted data (personally identifying information/HIPPA etc.) it will can and will be made freely available.

4. Does not state how long the researcher will make the data available.

5. Does not state how people will be able to access the data should the researcher change institutions or pass away.

Links to Other Resources

University of Nebraska-Lincoln, University Libraries Data Management Overview and Services

http://libraries.unl.edu/data-management

UNL's Data Management LibGuide

This guide contains detailed examples, explanations of terms, and links to resources on data management.

http://unl.libguides.com/datamanagement

O'Reilly Strata is an industry-centric website that focuses on data management, curation and practical applications. They offer free webinars, free and for cost publications, and a series of very prominent conferences focused on a variety of topics. They have a mailing list that you can join for an easy way to keep the industry perspective on bid data and data management in mind.

BigData by IEEE an annual conference collocated with several other IEEE conferences. The themes, keynotes, and schedule of events can provide a snapshot of what the organizers think are hot topics and trends in big data. (IEEE is a professional association dedicated to "advancing innovation and technological excellence for the benefit of humanity").

Data Conservancy is located at the Sheridan Libraries and Johns Hopkins University and is a nationwide collaborative effort. "The common theme of Data Conservancy's efforts is recognizing the need for institutional and community solutions to digital research data collection, curation and preservation challenges. DC tools and services incentivize scientists and researchers to participate in these data curation efforts by adding value to existing data and allowing the full potential of data integration and discovery to be realized.

Mailing Lists	**Twitter Hashtags and Handles**
• Esip-preserve • GEONET • PAMNET • Rdap • SLA-DST	• #opendata • #bigdata • #bigdatamgmt • #bigdataprivacy • Figshare • Research Data Summit • Open Science • Digital Science

Readings:

Scientific Data http://www.nature.com/sdata/ is an open-access, online-only publication for descriptions of scientifically valuable datasets, and exists to help you publish, discover and reuse research data.

Metadata for Digital Collections (How to do it manual) by Steven Miller. This is a basic introduction to metadata for those unfamiliar with how it works.

http://www.amazon.com/Metadata-Digital-Collections-How-To-Do-It-Librarians/dp/1555707467

Metadata Fundamentals for all librarians by Priscilla Caplan

http://www.amazon.com/Metadata-Fundamentals-Librarians-Priscilla-Caplan-ebook/dp/B004K6MCSI

New Content in Digital Repositories: The changing research landscape by Natasha Simons and Joanna Richardson

http://www.amazon.com/New-Content-Digital-Repositories-Professional/dp/1843347431

Other Data Initiatives to Follow:

COAR (Confederates of Open Access Repositories) has the goal of providing greater visibility and application of research through global networks of Open Access repositories - https://www.coar-repositories.org/

CHORUS (Clearinghouse for the open research of the United States) is a publisher initiated group for greater visibility and application of research through global networks of Open Access repositories - http://chorusaccess.org/

SHARE (SHared Access Research Ecosystem) is a higher education and research community initiative to ensure the preservation of, access to, and reuse of research outputs - https://sharewg.atlassian.net/wiki/pages/viewpage.action?pageId=8683527