Rubric for assessment of NSF data management plans

A product of IMLS National Leadership Grant LG-07-13-0328, "Analysis of data management plans as a means to inform and empower academic librarians in providing research data support"

Amanda L. Whitmire, Jake Carlson, Patricia M. Hswe, Susan Wells Parham, Elizabeth Rolando, and Brian Westra

This rubric is the **primary output of a** three-year National Leadership Grant for Libraries Demonstration Project to facilitate a multi-university study of faculty data management plans (DMPs). The purpose of creating an analytic rubric is to standardize the review of data management plans as a means to inform targeted expansion or development of research data services at academic libraries. Please see the supporting documentation for more detailed instructions on how to use this rubric. All documents related to this project are available at: *Whitmire, Amanda L, Jake Carlson, Brian Westra, Patricia Hswe, and Susan Parham. 2016. "The DART Project: Using Data Management Plans as a Research Tool." Open Science Framework.* http://osf.io/kh2y6.

This rubric is based on the generic <u>NSF DMP guidance</u>[1] and division- and directorate-specific guidance where extant. The general guidelines are:

1. The types of data, samples, physical collections, software, curriculum materials, and other materials to be produced in the course of the project;
2. The standards to be used for data and metadata format and content (where existing standards are absent or deemed inadequate, this should be documented along with any proposed solutions or remedies);
3. Policies for access and sharing including provisions for appropriate protection of privacy, confidentiality, security, intellectual property, or other rights or requirements;
4. Policies and provisions for re-use, re-distribution, and the production of derivatives; and
5. Plans for archiving data, samples, and other research products, and for preservation of access to them

Table of Contents

Version: 1.0 | Date: 23 March 2017 | Contact: Amanda Whitmire at thalassa@stanford.edu

---

[1] http://www.nsf.gov/pubs/policydocs/pappguide/nsf13001/gpg_2.jsp#dmp

Section 1: Types of data produced

Types of data, samples, physical collections, software, curriculum materials, and other materials to be produced in the course of the project

			Performance Level			
		Performance Criteria	Complete/detailed	Addressed issue, but incomplete	Did not address	Directorates w/this req.
General assessment criteria	**1.1**	*Describes what **types** of data will be captured, created, or collected*	Clearly defines data type(s). *E.g. text, spreadsheets, images, 3D models, software, audio files, video files, reports, surveys, patient records, samples, final or intermediate numerical results from theoretical calculations, etc. Also defines data as: observational, experimental, simulation, model output or assimilation* *"This project will produce: observational data in binary format, which is then converted by proprietary instrumentation software into ASCII text; Matlab mat-files for working data during the analysis phase; CSV and/or HDF5 files for data sharing and preservation."*	Some details about data types are included, but DMP is missing details or wouldn't be well understood by someone outside of the project *"We will collect zooplankton data from net tows. We will also collect elemental ratio data from preserved samples."*	No information about data types is included; fails to adequately describe data types.	All
Directorate- or division-specific assessment criteria	**1.2**	*Describes how data will be **collected, captured, or created** (whether new observations, results from models, reuse of other data, etc.)*	Clearly defines how data will be collected, captured or created, including methods, instruments, software, or infrastructure where relevant. *"Ride data will be collected via the smartphone apps, including both user entered information, such as observed road conditions or photos, as well as information captured automatically by the phone itself, such as GPS location."*	Missing some details regarding how some of the data will be produced, makes assumptions about reviewer knowledge of methods or practices. *"Data will be captured by sensors in the field"*	Does not include information regarding how the data will be collected, captured or created.	GEO AGS, GEO EAR SGP, MPS AST
	1.3	*Identifies **how much** data (volume) will be produced*	Amount of expected data (MB, GB, TB, etc.) is clearly specified. *"We expect to produce 5 - 7 GB of data."*	Amount of expected data (GB, TB, etc.) is vaguely specified. *"A small dataset will be generated."*	Amount of expected data (GB, TB, etc.) is not specified.	GEO EAR SGP, GEO AGS

Section 2: Standards for data and metadata

The standards to be used for data and metadata format and content (where existing standards are absent or deemed inadequate, this should be documented along with any proposed solutions or remedies)

		Performance Criteria	Performance Level			Directorates w/this req.
			Complete/detailed	**Addressed issue, but incomplete**	**Did not address**	
General assessment criteria	**2.1**	*Identifies **metadata standards** and/or metadata formats that will used for the proposed project*	The metadata standard that will be followed is clearly stated and described. If no disciplinary standard exists, a project-specific approach is clearly described. *"Data will be described using Darwin Core Archive metadata, and accompanied by readme.txt files providing information on field methods and procedures."*	The metadata standard that will be followed is vaguely stated. If no disciplinary standard exists, a project-specific approach is vaguely described. *"We will observe common data format and content protocols. For behavioral and the neuroimaging data, the participant log/response files include metadata embedded within the initial data collection. In addition, the behavioral and neuroimaging protocol/stimuli delivery files discussed above document the data and metadata format and content."*	The metadata standard that will be followed is not stated and no project-specific approach is described.	All
	2.2	*Describes **data formats** created or used during project*	Clearly describes file format standard(s) for the data. *"Data will be collected in plain text, Excel, SPSS, R, Matlab, Access DB, ESRI Shapefile, TIFF, JPEG, WAV, MP4, XML, HTML, or other software-specific file formats."*	Describes some but not all file formats, or file format standards for the data. Where standards do not exist, does not propose how this will be addressed. *"We produce data in the form of photon-counting events in a given time interval. The data is transferred to a laboratory server and stored. We also collect digital images of optical spectra from a spectrometer and store them on the server. We also collect digital samples of voltages monitoring various optical signals and store them on the server. The data will be available as raw data files."*	Does not include information about data format standards.	All
Directorate- or division-specific assessment criteria	**2.3**	*Identifies **data formats** that will be used for **storing** data*	Clearly describes data formats that will be used for storing data and explains rationale or complicating factors. *"NMR data will be saved in proprietary format to preserve embedded information, and converted to JCAMP files for ease of access and in case the proprietary systems fail or become unavailable."*	Only partially describes data formats that will be used for storing data and/or the rationale or complicating factors. *"Complex data analysis will be done using either Origin 8.0 or Matlab software and will be saved in a format that is easily transferable."*	Does not describe data formats that will be used for storing data and does not explain rationale or complicating factors.	GEO AGS, MPS AST, MPS CHE
	2.4	*If the proposed project includes the use of **unusual** data **formats**, the plan discusses the proposed solution for converting data into **more accessible** formats*	Explains how the data will be converted to a more accessible format or otherwise made available to interested parties. In general, solutions and remedies should be provided. *"Still images from microscopes will be converted from proprietary formats to Open Microscopy Exchange format (OME-Tiff) for preservation and sharing."*	Vaguely explain[s] how the data may be converted to a more accessible format or otherwise made available to interested parties.	Does not explain how the data will be converted to a more accessible format or otherwise made available to interested parties.	MST AST, MPS CHE

Section 3: Policies for access and sharing

Policies for access and sharing including provisions for appropriate protection of privacy, confidentiality, security, intellectual property, or other rights or requirements

		Performance Level			
	Performance Criteria	**Complete/detailed**	**Addressed issue, but incomplete**	**Did not address**	**Directorates w/this req.**
3.1	*Provides details on **when** the data will be made publicly available*	Clearly specifies when the data will be made available to people outside of the project. *"Data generated from our research will be shared incrementally throughout the time period of the project, and the full dataset will be available no later than one year after the project ends."*	Verifies that the data will be made available outside of the project but does not identify when, such as a time frame (e.g., duration of the project, or for a period after the conclusion of the project). *"It will be the policy of the project to publish relevant findings expeditiously in the peer-reviewed literature."*	Does not specify when the data will be made available outside of the project.	All
3.2	*Describes **how** the data will be made publicly available*	Includes specific details on the means by which the data will be made available. E.g., this may include a publically accessible data repository or a description of how the researcher or a 3rd party will provide access *"The data we collect will be made publicly accessible via submission to our university's repository service, which has capability for data set ingest and review."*	Provides vague or limited information on how the data will be made available, or details about sharing can be inferred from the mention of a repository or archive that will be used for depositing the data. *"The data will be made accessible to others upon request."*	Includes no details on the means by which the data will be made available.	All
3.3	***If** the data are deemed to be of a "**sensitive**" nature, describes what protections will be put into place to protect **privacy** or **confidentiality** of research subjects*	Actions that will be taken to address the sharing of sensitive data are clearly described *"Because of confidentiality issues, each subject will be assigned an arbitrary code. Personal information will be permanently removed prior to data analysis. All data will be stored and backed-up by the PI. The de-identified electronic data will be preserved on DVDs and external hard drives. Copies of these data will also be preserved offsite at a university storage facility. Completed surveys, consent forms, and written analyses of students' lessons will be stored in a locked file cabinet accessible only to the PI."*	Actions that will be taken to address the sharing of sensitive data are vaguely described. *"Prior to publication, the data will be maintained confidential and stored in the lab, and in the PIs machine."*	Actions that will be taken to address the sharing of sensitive data are not described.	All
3.4	*Describes what **intellectual property** rights to the data and supporting materials will be given to the public and which will be retained by project personnel (if any)*	Clearly defines the IP rights the public (or designated group) has in accessing the data and the rights retained by project personnel (if any). *"We do not anticipate that significant intellectual property issues involved with these data will arise. However, in the event that discoveries or inventions are made in direct connection with these data, access to the data will be granted*	Vaguely defines the IP rights the public (or designated group) has in accessing the data or that are retained by project personnel. *"Policies to access the data in these different formats are regulated by*	Does not address IP rights for the public, intended audiences or personnel in the research group.	All

General assessment criteria

			Clearly	Vaguely	Does not	
			upon request once appropriate invention disclosures and/or provisional patent filings are made."	*the journal or book copyrights."*		
	3.5	*Describes **security** measures that will be in place to protect the data from unauthorized access*	Clearly describes the security measures that will be put into place to prevent authorized access to the data. *"Personally identifying information (PII) will be removed and stored separately from the data files. Access to these separately stored PII files will have the added protection of encryption, such as via a password."*	Vaguely describes the security measures that will be put into place to prevent authorized access to the data. *"The PI will ensure restricted access to data by storing them on an external hard drive."*	Does not describe the security measures that will be put into place to prevent authorized access to the data.	All
Directorate- or division-specific assessment criteria	**3.6**	*If there are **factors that limit the ability to share data**, e.g. commercialization or proprietary nature of the data, plan describes those conditions and describes to whom the data will be made available and under what conditions*	Clearly defines the population to whom the data will be made available, as well as any conditions on access, with a justification for its limited release. *"Data will be made available to researchers on the condition of signing a non-disclosure agreement, to protect the confidentiality of the data."*	Vaguely discusses who will have access to the data or conditions on access. *"Data will be made available to researchers outside of this project on a case-by-case basis."*	Does not state who will be able to gain access to the data.	EHR; ENG; GEO AGS; GEO OCE
	3.7	*Describes **how long** the data will be retained and made available to people outside of the project*	Clearly defines the period of time during which the data will be retained and shared. *"There is no period of exclusive use by the data collectors. Users can access documentation and final monthly CO2 data files in perpetuity via the Scripps CO2 Program website."*	Vaguely describes the period of time during which the data will be retained and shared. *"We will deposit data files generated from our study into the institutional repository at the end of the project."*	Does not describe the period of time during which the data will be retained and shared.	CISE; EHR; ENG GEO AGS; MPS AST, MPS PHY; SBE
	3.8	***IF** data are going to be submitted to **supplementary materials** sections of peer-reviewed **journals**, the plan describes this*	Clearly indicates that data will be submitted as supplementary materials to peer reviewed journals *"We will submit CIF files for structures reported on in publications as supplementary files to the publisher, as required by the Journal of Organic Chemistry."*	Vaguely discusses plans to submit data as supplementary materials *"We will be sure to observe publishers' requirements to submit our data sets as supplementary materials."*	Does not discuss limiting factors, despite indications elsewhere in the plan that there are limiting factors in managing or sharing data.	MPS CHE
	3.9	*Describes data **types or formats** that will be used for **making data available***	Clearly identifies data formats or types that will be used for making data available. *E.g., all data will be shared vs. only a subset of raw data; quantitative, qualitative, observational, etc.* *"All NMR data will be converted to JCAMP format for sharing, and crystallography files will be shared as .cif files."*	Vaguely discusses plans to submit data as supplementary materials *"A complete set of all data will be archived and partially shared from the University Library Digital Archives."*	Does not identify any data formats related to data sharing.	CISE, ENG, EHR, SBE, BIO, GEO OCE, GEO EAR SGP, GEO AGS, MPS AST, MPS CHE

Section 4: Policies and provisions for re-use and redistribution

Policies and provisions for re-use, re-distribution, and the production of derivatives

		Performance Criteria	Performance Level			Directorates w/this req.
			Complete/detailed	Addressed issue, but incomplete	Did not address	
General assessment criteria	4.1	*Describes the policies in place governing the use and* **reuse** *of the data*	Clearly explains the policies in place governing future reuse of the data. *"The images used in a study were taken by a third party and licensed under CC-BY-SA, so any future uses would need to include attribution to the creators and use the appropriate license."*	Provides a general overview of how data may or may not be reused, or the applicability of the policy can be inferred from general/ broad/ blanket statements about data being made open or being kept private, or policies can be inferred based on the sharing location. *"The data will be shared in an unrestricted manner."* *"The data will be shared via Dryad."*	**Does not** explain the policies in place governing future reuse of the data.	All
	4.2	*Describes the policies for* **redistribution** *of the data*	Clearly explains policies or guidelines for future redistributions of data. *"In order to ensure proper credit is given, any future redistribution of the data should include attribution to the original creators and a link back to the original dataset"*	Provides a general overview of how and when data may be redistributed, or policies can be inferred based on the sharing location. *"The data will be shared via the NCEI database."* *"The data will be archived and shared in Dryad."*	**Does not** address redistribution of the data.	All
	4.3	*Describes policies for building off of the data, such as through the creation of* **derivatives**	Clearly describes guidelines or policies governing the creation of derivatives from the data generated by the project. *"The data will be shared under a CC-BY-NC-ND license"* *"No restrictions on re-use, distribution, or production of derivatives will be placed on the samples or the data following initial publication."*	Provides a general overview of the creation of derivatives from project data, or policies can be inferred based on the sharing location. *"Because the dataset contains information collected on human subjects, which could - when combined with other data - identify the original subjects, a policy explicitly forbidding that combination of data may be necessary."*	**Does not** discuss production of derivatives from the data.	All
Directorate- or division-specific assessment criteria	4.4	*Describes methods for* **communicating** *policies or guidelines for reuse, redistribution, and creation of derivatives*	Clearly states how the project/PI will disseminate information regarding the reuse, redistribution, or creation of derivatives. *Shared datasets will be accompanied by publically accessible metadata that will include all relevant reuse rights and restrictions. As the data from this project will be preserved and made accessible through a repository, the repository record will include this information.*	Refers to policies for various reuse and redistribution scenarios but does not account for how these policies will be communicated to other researchers.	**Does not** state how the project/PI will disseminate information regarding the reuse, redistribution, or creation of derivatives.	GEO AGS, MPS AST, MPS CHE

Section 5: Plans for data archiving and preservation of access

Plans for archiving data, samples, and other research products, and for preservation of access to them

		Performance Criteria	Performance Level			Directorates w/this req.
			Complete/detailed	**Addressed issue, but incomplete**	**Did not address**	
General assessment criteria	**5.1**	*Provides details on how the data will be archived*	Clearly indicates whether or not data will be archived, including digital data and physical samples where applicable *"All records, including unpublished data, will be maintained for a minimum of three years after publication or conclusion of the work. These digital files will be maintained in three locations to ensure redundant access: (1) on the lab/computer equipment where the data were originally collected, (2) on desktop computer(s) on which the data were analyzed/processed and (3) backed up on our long term data storage server housed on the campus of the University."*	The researcher's intent or commitment on archiving data is implied or indirectly stated but is not clear overall. *E.g., indicates that digital or physical data will be archived but isn't explicit about both.* *" Archiving and preservation of access to products from the proposed research is provided by and according to the policies of the University of Michigan system in connection with / Google™ and NSF-sponsored XSEDE™."*	The researcher either did not provide any information, or gives insufficient information for the reviewer to determine if the data will be archived or not. *"The researchers will keep copies of the data, but the data cannot be made available for replication."*	All
Directorate- or division-specific assessment criteria	**5.2**	*Describes how access to the archived data will be maintained*	Clearly describes how access to the archived data will be maintained over the long-term. *"The Data Dryad repository will enable long term access to the vegetation data, through their discovery interface which enables browsing and searching for data by multiple facets. Dryad will assign the data a DOI enabling it to be cited in publications."*	Vaguely describes how access to the archived data will be maintained over the long-term, or this can be inferred because the plan indicates that data will be deposited with a repository or archive. *"We will make the individual galaxy brightness profile dataset freely available as multi-dimensional FITS files as supplementary data to journal articles."*	**Does not** describe how access to the archived data will be maintained over the long-term. *"The project team commits to maintaining data archived in the repository for a period of at least three years following the project end date."*	CISE, EHR, GEO AGS, GEO EAR, MPS AST, MPS CHE
	5.3	*Describes plans for archiving and preserving digital data*	Clearly describes physical and cyber resources and facilities that will be used for the effective preservation and storage of research data. This might overlap considerably with information regarding how data will be shared. *"Data will be stored on departmental servers as they are being developed. Data will be transferred to the Dryad repository upon publication of our research and will be archived in accordance with Dryad's preservation policy: http://wiki.datadryad.org/Preservation_Policy"*	Vaguely describes physical and cyber resources and facilities that will be used for the effective preservation and storage of research data. *"Both the [participating universities] support campus-wide data storage plans under the auspices of their libraries. We will explore if these facilities can be used as clearing houses for our raw data and data products during this grant period, and, of course, as long-term data archives for these datasets independent of our personal data storage devices."*	**Does not** describe physical and cyber resources and facilities that will be used for the effective preservation and storage of research data. *"Research results will be disseminated in publications (book, review paper) for general Scientific audiences and selected data resources from case studies presented in the publications will be featured on an open-access website for students and interested colleagues."*	BIO, CISE, EHR, ENG, GEO AGS, GEO EAR, GEO OCE, MPS AST, MPS CHE, MPS DMR, MPS PHY, SBE

5.4	*Describes **plans** for archiving and preserving **physical** data*	Clearly describes what physical data will be archived, how much, where, and the conditions for archiving, including any safeguards against disrepair and damage. *"After data are extracted from the processed Vegetative matter, the collected samples will be freeze dried, inventoried applying metadata to connect it to the derived data, and stored in refrigerators maintained by the lab."*	Vaguely describes what physical data will be archived, how much, where, and the conditions for archiving, including any safeguards against disrepair and damage. *"The samples will be stored for a period of three years within the PIs laboratory post-publication."*	**Does not** describe what physical data will be archived, how much, where, and the conditions for archiving, including any safeguards against disrepair and damage.	ENG, GEO EAR, GEO AGS, GEO OCE, MPS CHE
5.5	*Identifies a **timeframe** for how long data will be archived*	Identifies clear timelines for length of retention, preservation and access. *"The digital data will be preserved in accordance with Dryad's preservation policy. Dryad does not specify a specific timeframe but states that for data in preferred formats Dryad will make "best effort(s) to ensure the full functionality of these files into the future, including format transformation/conversion as needed"*	Vaguely describes the period of time for which the data will be archived, or this can be inferred because the plan indicates that data will be deposited with a repository or archive. *"All data from this project will be made available for a period of at least 3 years after the end date of the project. Since the data from this project will be hosted on the UM website, it will likely be available long after the 3 year period specified above is over."*	**Does not** identify clear timelines for length of retention, preservation and access.	CISE, EHR, ENG, SBE, MPS_PHY, MPS_CHE, MPS_DMR, GEO_AGS, MPS_AST, GEO_EAR_SGP
5.6	*Plan discusses the types or formats of data the investigator expects to **retain** in their possession*	If not all data are made publicly available, plan clearly describes the types or formats of data to be retained (e.g., all data will be shared vs. only a subset of raw data; quantitative, qualitative, observational, etc.). *"The PI will retain raw data from the project in case the publicly available processed data needs to be reprocessed or validated." or "While CSV files will be shared openly via the IR, the PI will retain proprietary file types for internal ease of use and for use by others upon request."*	Provides vague/limited details regarding the types of data that will be retained. *"Our project will ensure that the research data are migrated to new formats, platforms, and storage media as required by good practice in a secure environment (e.g., lockable computer systems with passwords, firewall system in place, power surge protection, virus/malicious intruder protection)"*	Provides no details regarding the types of data that will be retained.	CISE, ENG, SBE