Reviewer's Worksheet for NSF Data Management Plans

The table & checklists cover NSF's requested components of the proposal's data management plan. A ☆indicates details found in more thorough plans, and a quick measure of quality when checked. See pg.2 for more examples and guidelines.

Research product	Source	For	mat	Size	Preserved	Shared	
computer code, curriculum items, physical samples Pl'	ta repository, strument, interviews, s prior project	JPG, MA Excel tal device's	ble,	>1TB, 20K files	(how?) Discarded, PI retains, data archive	(how?) By request, website, repository	
;							
Data Sharing \underline{i} $\underline{e}\underline{g}$.			Data n	nanageme	ent during projec	t:	
☐ Is data publically accessible?			☐ Storage: has a backup plan <u>i. eg.</u>				
☐ When will data be shared?				ocation &	media used:		
☐ Who administers?				2+ copie	s with 1 off-site		
$\square \ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \$			☐ ☆ Specifies who is responsible				
Preparation of data for sharing: <u>i</u> eg. ☐ Uses their research field's metadata standards			☐ ☆ Data security / access controls <u>i</u> eg				
☐ AND/OR creates description	n sufficient for re-us	e		organizii	ng files	<u>i</u> <u>eg</u>	
Metadata or supplementary files explaining:			☐ ☆ Version control				
content/ file structure/ procedures/ codebook or			□ ☆ Collaboration coordination				
variable-level detail			Data retention after the project: \underline{i}				
□ ☆ Metadata associated with digital files			☐ Where is data preserved?				
☐ ☆ Will convert files to non-proprietary formats			☐ How long?				
Data sharing policy: <u>i</u> <u>eg</u> .			☐ Who administers?				
Gives conditions for re-use Accounts for:			☐ ☆ Gives reasons for preserving data (especially raw data)				
	fiors)/socurity issues			Heing an	archivo corvico o	\r	
 □ ☆ privacy (personal identifiers)/security issues □ ☆ intellectual property (copyrights, patents) 			☐ ☆ Using an archive service or repository?				
☐ If a plan states there is no	data to manage or sh	are, hav	e they ju	stified it?	<u>i</u>		
ervices of archive (if specified fo	r preservation and/o	r sharing	data) <u>i</u>				
Archive type	Preservation	Preservation activities			Data sharing services		
PI's Institutional repository (documents)	☐ Data integrit	☐ Data integrity checking			☐ Public access to data files		
☐ ☆ Research data repositor	☐ ☆ Migration to new formats, media ☐ Persistent data citation			a citation			

Reviewer's Worksheet for NSF Data Management Plans

p. 2

These guidelines and examples correspond to the checklist's table of data products and other sections.

Research Product table guidelines

Component	Guidelines	Examples	
Research product	NSF considers research products to include digital data, physical samples or collections, and supporting materials. Pl's should list primary data used to support results, and preliminary or "raw" data acquired for analysis. Both types require management and storage during the project, even if only primary data are archived and shared.	 Types of data: observations, experimental results, models, curricular items, lab notebooks, specimen or material samples Digital data: tables, images, computer code, video, documents Organized as: spreadsheet, database, file directories 	
Source	Indicating data sources, such as the instrument or collection approach, is optional but gives a sense of the research workflow, and is particularly relevant for indicating data from prior work, project collaborators, or source data not owned by PI.	 Data repository (named) Instrument, analytical software, field collection Co-PI's experimental results from her lab at X University Dataset created from PI's prior research 	
Format	Digital file formats in which data will be stored and shared should be listed as file types (images/video/tables) and/or specific software formats. If latter are proprietary, or customized for a research instrument, a thorough plan states whether shared data will be converted to a commonly accessible format.	 JPG image, FITS image MATLAB MS Excel table, converted to CSV SQL database Instrument's proprietary format, converted to spreadsheet tables 	
Size/scope	Estimating size and number of files gives a sense of scale of data managed, stored and shared.	■ >1TB, ~2000 images ■ 200MB of documents.	
Preserved (how?)	Directorates or PI's institution may ask for at least 3-5 years preservation post-project. PI's should indicate which of the research products listed here will be preserved (most commonly, the primary data), the method of preservation and who will manage it.	 Raw data will be discarded. PI retains two DVD copies checked annually for file integrity. Data for publications will be archived with Dryad (details given) 	
Shared (how?)	If the PI is committing to sharing certain data products, the method of distribution should also be indicated. Direct contact with the project PI? Access via a personal or team website? Access via an archive, repository, or data center for their field? Sharing data through archives increases access and discoverability, and may aid in preservation of datasets (c.f. guidelines below for Services of a Data Archive).	 Data will be released by direct request to the PI. The PI will share published data on her departmental webpage. Archive/Repository: "Gene sequences shared through GenBank", "demographic analysis on ICPSR" "Neuro imagery processing algorithms on NITRC" 	

Reviewer's Worksheet for NSF Data Management Plans Sharing data

p. 3

- Shared or Accessible Data: NSF expects data sharing to follow the norms of the PI's research community, but encourages efforts to broaden the audience and the range of data shared. Data can often be of unanticipated interest in the future if it can be located, understood, and cited. The DMP should ideally indicate the value of the shared data for specific audiences. "Sharing" can include direct release to interested parties upon request. "Accessible" generally means unmediated distribution of data through an online resource or database.
 - Example: "All data types will be made publically accessible after de-identification and an embargo period of one-year post grant completion and first publications. Raw MRI data will be shared through the PI's university data repository and will use a unique durable identifier to assure that the dataset and associated metadata can be appropriately cited."
- Preparation of Data for Sharing: Datasets intended for use by others require data collections and files to be associated with sufficient contextual documentation, also called metadata. Some research fields have established metadata standards for files, such as Content Standard for Digital Geospatial Metadata (CSDGM), Climate and Forecast (CF) metadata convention, or Chemical Markup Language (ChemML). Pls should also produce their own descriptive metadata for shared datasets, particularly if standards are uncommon for their field. Metadata may take the form of "readme files" that explain variables and file structures. Machine readable metadata, such as XML are preferable for online access.
 - Example: "At the beginning of research, the PI will develop a convention for storing metadata on files. This metadata will include: sampling procedures, the data source, and geographic information following the Content Standard for Digital Geospatial Metadata (CSDGM) standard, along with a list of abbreviations and other technical details. The metadata will be generated manually each time a file is produced, and support XML encoding."
- **Data Sharing Policies:** This section should describe policies for the level of access given to users, limitations and protections for data shared, and constraints on the requester's ability to re-use and redistribute the data.
 - Example: "Data use by other investigators in subsequent work requires citation and a statement acknowledging the PI's lab for collecting the data and NSF for funding the project. Publically accessible data will be de-identified to maintain confidentiality and follow HIPAA regulations. No other restrictions shall be imposed on data use."
- Plans stating they have no data to share, preserve, or manage during research should be accompanied by a justification: NSF's Data Sharing Policy states: "Investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants." If a plan claims that no data can be shared, has the PI accounted for reasonable options for disseminating some products of research? For example, if personal identifiers of human subjects limit direct sharing of the data, can identifiers be removed for a sample set or case examples? Are there curricular materials to share? If there are intellectual property restrictions claimed, can materials be released at a future date, following patents? If PIs claim data will have no future use to others or themselves, is it relevant to maintain data associated with publications, including possible online links to source datasets? The extent of efforts to share data may be a factor to consider in assessing the quality of the plan.

Reviewer's Worksheet for NSF Data Management Plans Data management during project

p. 4

- **Storage & Backup:** Although not all directorates require discussion of storage, a DMP should ideally mention a backup plan. An exemplary plan regularly creates at least 2 copies, with one off site. Backups to institutional and "cloud" servers should be supplemented by media the PIs or their project or institution directly controls, such as hard drives or DVD.
 - **Example:** "All digital audio files of interviews and transcripts will be backed up, daily, to an encrypted hard-drive, and uploaded to the Co-PI's account on the university's password-accessed server, which will serve as the offsite backup for the field project."
- **Data Security & access controls:** A good plan considers security against tampering, even for data intended for public access. Access to data with *identifiable* human subjects should be adequately protected, including passwords, encrypted files, or locked physical media. Password-protected PCs and servers are generally adequate for non-sensitive data.
 - **Example:** "As far as possible, all material will be de-identified using a coding key which will be kept for the PI's and co-PI's reference on a password protected hard drive secured in a locked cabinet."
- Conventions for naming, file organization, version control & collaboration: NSF Directorates do not require discussion of file organization or file naming conventions; however, a thorough plan will indicate atypical aspects of organization, such as the use of version controls, building a shared database, or other protocols for coordination among multi-institutional collaborators. Good file organization aids use of data during the project and the longer-term sharing of data.
 - Example: "Naming conventions will be developed to facilitate archiving and retrieval of data, based on the date of acquisition and the type of data referenced. Source data files will be stored along with associated derived data files. Associated file properties will document provenance of derived datasets and synthesis protocols/parameters for the creation of project materials."

Data retention after the project

Preservation and Archiving (vs. Storage): Preserving or archiving data is a more involved process than storage during a project. Preserving digital data should include periodic integrity checks, restoring degraded media, upgrading data formats, and maintaining adequate documentation of content. Archiving, particularly in a data repository, encompasses both active preservation of the digital object and increased discoverability and access to those data. The DMP should indicate which data are chosen for preservation, for how long and who is responsible, the PI or a data archive. Does the plan justify keeping preliminary or raw data?

Services of data archive: (if specified for preservation and/or sharing data) DMPs indicating they will deposit to a data archive, repository, or data center are agreeing to a notable effort in meeting NSF's data sharing and preservation policy, especially if data repositories are uncommon in the PI's field. The plan should name the repository. Specifying its services is not essential, but would illustrate the quality of access the project will provide. Some schools open their institutional document repository for data files. Disciplinary repositories and other archives designed for research data often provide additional services, such as the creation of persistent, unique identifiers for citation, format migration, disaster recovery plans, and an interface for online data access. Free public access is preferable for NSF projects.

Version 20141026