UDC Guidelines on Data Management Plans

The University of the District of Columbia's (UDC) mission as a public land-grant institution of higher education requires us to engage in research of importance to the District of Columbia and the nation, to ensure the integrity and reliability of the research, and to facilitate transfer of the results of the research to benefit society and expand the knowledge base of the discipline. The University's policy is that the Principal Investigator (PI) is responsible for understanding and fully implementing the specific sponsor's requirements in all respects. When research is supported by public funds, the data obtained should be appropriately and responsibly managed, and they should be shared to maximize its benefits to the taxpayer.

The National Science Foundation requires proposals to include plans for managing data and sharing of the results of research it funds. FastLane no longer permits submission of any proposal that is missing a Data Management Plan. The Data Management Plan cannot exceed two pages in length, and it will be reviewed as part of the intellectual merit or broader impacts of the proposal, or both, as appropriate. The Data Management Plan should detail how the proposal will conform to NSF policy on the dissemination and sharing of research results (see AAG Chapter VI.D.4). If no data are involved in a particular proposed project, the Data Management Plan can be very short and simply state this fact.

Many factors—some of them discipline- or field-specific—are involved in the management of data. The Data Management Plan should reflect best practices in the area of research, and it should be appropriate to the data produced. The process of preparing a Data Management Plan gives the PI and collaborators an opportunity to address prior to starting the project, such matters as:

- The types of data that the project might generate and eventually share with others, and under what conditions;
- How data are to be organized, managed, maintained, archived, curated, and protected against distortion;
- Factors that might complicate or compromise the data or their management, for example possible legal or ethical restrictions, human subjects concerns, etc.;
- The level at which data are to be aggregated, prior to sharing them with others in the scientific community, given that community's norms on data;
- The mechanism for sharing data and/or making them accessible to others;

- Other types of information that should be maintained and shared regarding data, e.g. the way it was generated, analytical or procedural information, and any associated metadata.

What type of data is covered under a Data Management Plan? The federal definition of data covered by a Data Management Plan is provided in 2 CFR 215, *Uniform Administrative Requirements for Grants and Agreements with Institutions of Higher Education, Hospitals, and Other Non-Profit Organizations* (also known as OMB Circular A-110). Research data are defined as:

> "the recorded factual material commonly accepted in the scientific community as necessary to validate research findings, but not any of the following: preliminary analyses, drafts of scientific papers, plans for future research, peer reviews, or communications with colleagues. This "recorded" material excludes physical objects (e.g., laboratory samples). Research data also do not include:
>
> > (A) Trade secrets, commercial information, materials necessary to be held confidential by a researcher until they are published, or similar information which is protected under law; and
> >
> > (B) Personnel and medical information and similar information the disclosure of which would constitute a clearly unwarranted invasion of personal privacy, such as information that could be used to identify a particular person in a research study."

Sample Outline for a Data Management Plan: A Data Management Plan typically contains the following sections; however please check your sponsor's specific requirements:

- Data description and nature of the data: This section should provide a general overview of the nature of the data or other materials produced under the NSF-sponsored project. What are the characteristics of the data? What type of data will be generated? If your data are of a sensitive nature (related to human subjects for example), it should be noted in this section and addressed more fully in later sections.

- Standards to be used for data and metadata format and content: This section should provide short summary of the data standards and metadata standards you will use over the course of your projects. The term, "metadata" refers literally to "data about the data," and they usually take the form of a list of elements used to describe the data. What file formats will be use for the data (if applicable)? What metadata will be collected and maintained to make the data meaningful?

How will the project save the details of the data and metadata? How will the accuracy and validity of the data and metadata be assured?

- Describe the method for preserving the data: This section should provide a short summary of how the data will be preserved and made available for sharing. Some of the issues to be addressed are:
  - Will the data be shareable? (If the data will not be shareable, please explain why here.)
  - How and when will you make the data available? (If there are any embargo periods due to contractual arrangements please detail those here.)
  - What is the process others would use for gaining access to the data?
  - Does the original data collector/ creator/ principal investigator retain the right to use the data before opening it up to wider use? If yes, for how long?
  - What, if any, provisions will be made for appropriate protection of privacy, confidentiality, security, intellectual property, or other rights or requirements if necessary?

- State how long the data will be kept: 2 CFR 215 (OMB Circular A-110) mandates that original data be kept at least 3 years from the end of the project providing there are no ongoing investigations. However, the common practice is to keep the original data in perpetuity. Some issues you might want to address in this section are:
  - What is the long-term strategy for maintaining and archiving the data?
  - Where will the data be stored?
  - What transformations will be needed to allow data sharing (de-identifying or aggregating the data, etc.)?
  - What metadata and/or documentation will be created and where will it be stored?
  - How long will the data be kept?

References and helpful links:

- NSF Overview of the Dissemination and Sharing of Research Results (including Directorate-level guidance): http://www.nsf.gov/bfa/dias/policy/dmp.jsp
- NSF Data Sharing Policy: http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/aag_6.jsp#VID4
- NSF Data Management Plan Requirements: http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/gpg_2.jsp#dmp
- NSF Data Management Plan “Frequently Asked Questions”: http://www.nsf.gov/bfa/dias/policy/dmpfaqs.jsp

Sample data management plans[1]

1. Sample Data Management Plan: Physical Sciences and Engineering

The data obtained during the proposed project will consist of XXXXX, as described in the main body of the proposal. These records will consist chiefly of [*describe specific type of data being collected in the proposal]* obtained via [*describe the methodology used to create or generate the data*.] These data will be recorded via computerized data acquisition software, with essential metadata present either as header in the relevant electronic files, or included along with the indexed laboratory notebook narrative.

These data will provide an experimental look at the [*subject matter]*, as delineated in the main body of the proposal. As such, they will be of interest to the [*insert the main community of interest here]* community, as well as to the [*if there are other scientific communities who might be interested list them here]* communities.

The [*DATA GENERATED FROM THIS PROJECT*] will be computer files generally in the form of tab-delimited numbers with header information. These computer files will be accompanied by dated laboratory notebooks, as well as by numerical data files analyzed using [*insert the type of software used here*.]

The electronic data will be preserved in multiple on-site backups in the form of DVDs and RAID hard drive storage. Copies of the electronic data will be preserved off-site at XXX storage facility. Original laboratory notebooks will be secured by the PI in [*laboratory, desk, file cabinet etc.]*. If requested, access to the data will be provided via contact with the PI. Data will, in principle, be available for access and sharing as soon as is reasonably possible, and not longer than two years after the acquisition of the data. The data will be preserved for at least three years beyond the award period, as required by NSF guidelines. This project will not involve the acquisition of either animal or human subject data.

We do not anticipate that there will be any significant intellectual property issues involved with the acquisition of the data. In the event that discoveries or inventions are made in direct connection with this data, access to the data will be granted upon request once appropriate invention disclosures and/or provisional patent filings are made.

The data acquired and preserved in the context of this proposal will be further governed by the University of the District of Columbia's policies pertaining to intellectual property, record retention, and data management.

2. Sample Data Management Plan: Biosciences

The data generated through the work described in this proposal will consist of [*nucleotide and protein sequences, graphs, three-dimensional structures, reconstructed images, photographs, hand-recorded observations, biological expression signatures, movies, etc*.*]* in both raw and curated forms, and will include relevant statistical analyses. These data will consist primarily of *[describe specific type of data being collected in the proposal, e.g., "arrays of gene expression data from cells grown in control and test conditions as described in the body of the proposal"].* These data will be collected using the instruments and methods described in the proposal and will include datasets generated from commonly accepted data acquisition software, with essential metadata presented as headers in the relevant electronic files, or included along with the indexed laboratory notebook narrative.

Records of results will be labeled and will be stored as either hard copy or as digitized images. [*If needed, add a statement such as, "In some cases, observations of [organism or cell]*

[1] Adapting from Rice University http://osr.rice.edu/forms/datamanagementPlans.pdf

behavior, viability, or phenotype will be retained in hand or electronic notation that will be dated and labeled in dated laboratory notebooks.]

These data will provide an experimental look at the *[describe project briefly, e.g., "responses of cells to experimentally manipulated changes in gene expression].* As such, they will be of interest to the *[in this case, genomics]* community, as well as the *[any other community, e.g., pharmacology]* community.

All members of the investigative team with access to data will receive instruction in the Responsible Conduct of Research (RCR); this instruction will include CITI training and additional training onsite. Original data notebooks will be retained in a secure location in the PI's laboratory, with electronic data backed up on The University of the District of Columbia's off-site storage servers whenever the nature of the data makes such archiving possible. If requested, data will be made available for sharing to qualified parties by the PI, so long as such a request does not compromise intellectual property interests, interfere with publication, invade subject privacy, betray confidentiality, or precede data curation. Data that are shared will include standards and notations needed to interpret the data, following commonly accepted practices in the field. Data will be available for access and sharing as soon as is reasonably possible, normally no longer than two years after its acquisition.

Data will be retained for at least three years beyond the award period, as required by NSF guidelines. In the event that discoveries or inventions are made in direct connection with this data, access will be granted upon request once appropriate invention disclosures and/or provisional patent filings are made. Key data relevant to the discovery will be preserved until all issues of intellectual property are resolved.

Insert one of the following statements, modified as necessary:
A) The data to be acquired in the proposed project will not involve human subjects.

B) The data to be acquired in the proposed project will include human subjects data that require Institutional Review Board approval. As detailed in the human subjects section of the proposal, all rules and regulations related to privacy (i.e., HIPPA) will be observed with specific regard to collected data.

> *If data are de-identified:* To preserve confidentiality, each subject will be assigned an arbitrary code that will be associated with the data, and only data stripped of all potential identifiers will be stored in the collected and curated data sets. One file that contains the links to subject names and identifiers will be kept in a password-controlled file that will be accessible only to the subject coordinator. De-identified electronic data will be stored on external hard drives and DVDs, and copies of de-identified data will be preserved off site on the XXX server. *[If needed, expand on this to cover details specific to the proposed project.]*
>
> *If data include identifiers:* It is recognized that potential identifiers may exist in the datasets. All hard copies of data that contain such identifiers will be kept in a locked area of the laboratory or office, with key access available only to the PI and subject coordinator. All electronic data will be encrypted and stored in password-protected files that will be accessible only by the PI and the subject coordinator. All members of the UDC community who may have access to the digitized storage files will receive appropriate HIPPA training from CITI.

C) (IF NECESSARY) The data to be acquired in the proposed project will include data collected from experimental animals, as described in the institutional animal care and use section of the proposal. All such data will be collected only under IACUC-approved protocols that will be

preserved with approval numbers and signatures by the PI. Observations of animal behavior, viability, or phenotype will be preserved as described above.

The data acquired and preserved in the context of this proposal will be further governed by The University of the District of Columbia's policies pertaining to intellectual property, record retention, and data management.

3. Sample Data Management Plan: Social and Behavioral Sciences

The proposed project will include human subjects data consisting of [*insert a description of the data to be collected and how it will be collected.]*

The [*type of data]* data will be used to determine XXX. The data will be used to verify XXX [*insert your hypothesis here*. The demographic data will be needed for published reports to convey the characteristics of the subject population.

The demographic data will be collected from a questionnaire administered by the PI and [*list others on the project who may do data collection]* associated with this project, and will be entered into [xxx]. The data will consist of [*describe your data here]*. Analyzed XXX data will be stored in the XXX file format.

Since these data will be from human subjects, approval for human subject research will be obtained through the UDC Institutional Review Board. As detailed in the human subjects section of the proposal, because of confidentiality issues, each subject will be assigned an arbitrary code, which will be associated with the XXX data. [*describe your data protocol as detailed in your IRB application].* Personal information (name, date of birth, etc.) will be permanently removed prior to data analysis.

The de-identified electronic data will be preserved on [*insert appropriate storage locations here.]* Copies of these data will also be preserved offsite at XXX storage facility. Completed questionnaires (as well as representative blank questionnaires) and human subject consent forms will be stored in a locked area of the PI's [*laboratory, desk, file cabinet etc.]* accessible only to the PI and [*whomever else will have access].*. If requested, access to the de-identified data will be provided [*by contacting the PI*.] [NOTE: If you will ultimately place your data in a data repository indicate that here.] Data will in principle be available for access and sharing as soon as is reasonably possible, normally not longer than one year after publication of the data. The data will be preserved for at least three years beyond the award period, as required by NSF guidelines.

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 upon request once appropriate invention disclosures and/or provisional patent filings are made.

The data acquired and preserved in the context of this proposal will be further governed by The University of the District of Columbia's policies pertaining to intellectual property, record retention, and data